Polysémie verbale et calcul du sens

Guillaume Jacquet

ECOLE DES HAUTES ETUDES EN SCIENCES SOCIALES

LaTTICe CNRS
Langues, Textes, Traitements informatiques, Cognition

Doctorat nouveau régime

Discipline : Sciences Cognitives

Guillaume **JACQUET**

POLYSEMIE VERBALE ET CALCUL DU SENS

Thèse dirigée par Bernard VICTORRI

Soutenue le mardi 6 décembre 2005

Jury :

M.	Jean PETITOT	Président
Mme	Adeline NAZARENKO	Rapporteure
M.	Benoît HABERT	Rapporteur
Mme	Frédérique SEGOND	Examinatrice
M.	Didier BOURIGAULT	Examinateur
M.	Bernard VICTORRI	Directeur de Thèse

A mon père et à ma fille…

Résumé

La désambiguïsation lexicale automatique fait partie des domaines du Traitement Automatique de la Langue (TAL) les plus productifs. Deux raisons expliquent cette productivité : d'une part l'intérêt considérable de la désambiguïsation lexicale automatique pour un nombre important d'applications, d'autre part l'absence de consensus sur la manière d'appréhender cette tâche.

Nous proposons un modèle d'un genre nouveau fondé sur la théorie de la construction dynamique du sens. Cette théorie donne une place centrale à la polysémie et propose une représentation géométrique du sens permettant de rendre compte de l'aspect continu du sens.

Nous développons ce modèle pour la désambiguïsation des verbes. Pour cela, nous proposons de donner une place nouvelle à la syntaxe en postulant que les constructions syntaxiques sont porteuses d'un sens intrinsèque. Parallèlement, nous proposons une méthode de calcul automatique de classes sémantiques pour les éléments lexicaux qui sont rattachés au verbe, ceci afin d'améliorer les performances du calcul de désambiguïsation.

Mots-clés

Traitement automatique des langues, désambiguïsation, sémantique, polysémie, construction syntaxique, grammaires constructionnelles, construction dynamique du sens, synonymie, classes sémantiques, espace sémantique, paraphrases, espace distributionnel, modélisation, information mutuelle.

Remerciements

Cette thèse correspond à l'aboutissement d'un objectif qui me semblait il y a encore quelques années absolument inaccessible. Elle concrétise de nombreuses années de réflexion et de travail, et j'ai pour cela de nombreuses personnes à remercier.

Je tiens en premier lieu à remercier Bernard Victorri pour avoir accepté de diriger cette thèse ainsi que le DEA qui l'a précédé. Pour sa détermination à être disponible en toutes circonstances, pour les nombreuses discussions qui m'ont éclairé, orienté, poussé à aller plus loin dans ma réflexion.

Jean Petitot pour avoir accepté de présider le jury.

Adeline Nazarenko et Benoît Habert pour avoir accepté d'être les rapporteurs de cette thèse ainsi que pour leurs commentaires très justes qui ont permis d'affiner mes travaux.

Frédérique Segond et Didier Bourigault pour avoir pris le temps d'examiner cette thèse.

L'équipe du laboratoire Lattice pour les réunions enrichissantes, les projets développés en communs et les discussions plus informelles mais toutes aussi utiles. Plus particulièrement Catherine Fuchs pour m'avoir accueilli dans son laboratoire, Sophie Prévost, Laure Sarda et Fabienne Venant pour les projets que nous avons montés en commun, leurs relectures toujours pertinentes et aussi pour avoir écouté mes états d'âme professionnels et personnels.

Didier Bourigault, Jacques François, Jean Luc Manguin et Cécile Frérot, membres du projet ILF « Polysémie verbale et constructions syntaxiques » avec qui les discussions et travaux partagés furent importants pour cette thèse.

Didier Bourigault une nouvelle fois pour sa disponibilité et sa générosité en corpus et autres bases textuelles.

Fabienne venant pour les très nombreuses discussions en tous genres, mais toujours bénéfiques.

Sylvie Bordin qui en échange de quelques assistances informatiques m'a simplifié de nombreuses taches administratives.

Jennifer Mercier, Yves Bestgen, Laurence Delort, Emmanuel Giguet, Nabil Abdelaoui, Sylvain Kahane.

Je remercie aussi l'ensemble de ma famille, particulièrement mon père et ma fille, Clémence, à qui je dédie cette thèse, mais aussi mes grands parents Jacqueline et Armand, ma mère, Caroline, Sophie, Fred, Marion, Annabel, Sylvie, Alizée, Jérôme, Julie, Eloïse, Léonie, Roselyne, Thierry, Stéphanie, Michel, Eric, Jacques, Martine, Jean-Claude, Gérard, Claire, Delphine, Simon, Paul, Edith, Sylvaine, Fanny, Perrine, Karine, Florie, Emilie, et j'en oublie sûrement… Tous ont suivi de près ou de loin le déroulement de cette thèse et je les remercie notamment pour leurs nombreux messages de soutien.

Mes amis Antoine, Matthieu, Pascal, Elodie, Paul, Eloïse, Babo, Nanne, Lionel, Laurent, Yann, Virginie, …

Enfin, je remercie tout particulièrement Marie qui a sans aucun doute le plus suivi et subit cette thèse. Je la remercie pour son amour, son soutien moral et gastronomique, ses conseils et sa compréhension.

Sommaire

Résumé 3

Table des figures 9

1. Introduction 11

1.1. Objectif de la thèse 11

1.2. Problématique 15

1.3. Plan de la thèse 20

2. La polysémie 22

2.1. Polysémie et homonymie : continuités et ruptures de sens 22

2.2. Phénomènes de parole et degrés de lexicalisation 25

2.3. Métonymies intégrées et coercition de type 28

2.4. Approches de la polysémie lexicale 31

2.5. Spécificités de la polysémie verbale 36

3. Désambiguïsation lexicale automatique 42

3.1. Six travaux représentatifs 45

3.1.1 Yarowsky (2000) 45

3.1.2 Peng Ito et Furugori (2001) 49

3.1.3 Wilks et Stevenson (1998) 54

3.1.4 Inkpen et Hirst (2003) 57

3.1.5 Resnik (1995) 62

3.1.6 Schütze (1998) 67

3.2. Evaluation et confrontation des modèles 71

3.2.1 Que peut-on comparer ? 72

3.2.2 La granularité de sens 73

3.2.3 Bornes inférieures et bornes supérieures 73

3.2.4 Senseval 74

3.3. Deux sources de connaissances essentielles : les relations syntaxiques et les classes sémantiques 79

4. Un modèle dynamique de calcul du sens 84

4.1. Continuité de sens 86

4.1.1 Monter 87

4.1.2 Compter 90

4.1.3 Jouer 91

4.2. Fondement théorique 93
4.3. Développement logiciel 95
4.3.1 Construction d'un espace sémantique 95
4.3.2 Espace sémantique du verbe compter *100*
4.3.3 Espace sémantique du verbe jouer *102*
4.3.4 Région sémantique associée à un synonyme 103
5. Une méthode de calcul du sens 106
5.1. Le modèle théorique de construction dynamique du sens 106
5.2. La méthode de calcul 112
5.2.1 Le principe : calculer un degré d'affinité entre une unité co-textuelle et une clique 112
5.2.2 Calcul du degré d'affinité : première méthode 114
5.2.3 Calcul du degré d'affinité : deuxième méthode 115
5.2.4 Calcul de la fonction potentielle 117
5.3. Les corpus 118
5.3.1 Frantext 119
5.3.2 Le journal Le Monde 120
6. Les constructions verbales : calcul de l'influence du co-texte lexical 121
6.1. Logiciel 121
6.2. Résultats : Région sémantique associée à un nom co-textuel 125
6.2.1 Monter *et le co-texte* escalier *126*
6.2.2 Monter *et le co-texte* diamant *130*
6.2.3 Monter *et le co-texte* projet *133*
6.3. Discussion 135
7. Les Grammaires Constructionnelles 137
7.1. Généralité sur les grammaires constructionnelles 137
7.2. La grammaire constructionnelle de Goldberg 140
7.2.1 Le modèle d'interaction entre verbes et constructions 141
7.2.2 La construction de mouvement induit (caused motion construction) 146
7.3. Discussions de Goldberg et des grammaires constructionnelles 149
7.3.1 Sens de base des constructions 149
7.3.2 Correspondance de rôles 150

7.3.3 La construction de mouvement induit ... *151*
7.3.4 Conclusion ... *153*
8. Les constructions verbales : calcul de l'influence du co-texte syntaxique ... 154
8.1. Méthode ... 155
8.2. Développement logiciel ... 160
8.3. Evaluation qualitative sur le verbe *jouer* ... 163
8.3.1 La construction prépositionnelle V sur SN ... *164*
8.3.2 La construction prépositionnelle V avec SN ... *166*
8.3.3 Les constructions V à SN *et* V de SN ... *168*
8.3.4 Discussion des résultats ... *171*
8.4. Evaluation psycholinguistique ... 172
8.4.1 Taux d'adéquation ... *173*
8.4.2 Résultats ... *174*
8.5. Combinaison lexique-syntaxe ... 179
8.6. Discussion ... 183
9. Les constructions verbales : améliorations du modèle ... 184
9.1. Degré « d'essentialité » d'un complément ... 186
9.1.1 Utilisation du Web comme « corpus » ... *188*
9.1.2 Méthode ... *190*
9.1.3 Test1 : fréquence de la construction verbe + complément ... *191*
9.1.4 Test2 : antéposition du complément ... *192*
9.1.5 Premiers résultats ... *192*
9.1.6 Discussion ... *197*
9.2. Des classes de sélection distributionnelle ... 199
9.2.1 Méthode ... *200*
9.2.2 Constructions des CSD (Classes de sélection distributionnelle) ... *202*
9.2.3 Evaluation des résultats ... *206*
9.2.4 Discussion ... *210*
10. Conclusion et perspectives ... 212
10.1. Améliorations méthodologiques ... 212
10.1.1 Points algorithmiques ... *213*
10.1.2 Les corpus ... *214*

10.1.3 Les constructions *216*
10.1.4 Combinaison lexique - syntaxe *217*
10.2. Applications 219
10.2.1 Paraphrases de constructions verbales *219*
10.2.2 Applications directes *221*
10.2.3 Retour à l'analyse syntaxique *223*
10.3. Au delà de la désambiguïsation des verbes 224
10.3.1 CSD et entités nommées *224*
10.3.2 CSD et désambiguïsation nominale *226*
10.4. Vers un modèle global de la désambiguïsation d'un énoncé 227
Annexes 229
A. Annexe 1 : visualisations 3D 229
B. Annexe 2 : extraits de code 232
C. Annexe 3 : schéma des variantes de la construction de « mouvement induit » de Goldberg (1995 : 164) 234
D. Annexe 4 : exemple de sortie de Syntex 235
E. Annexe 5 : extrait d'un questionnaire distribué pour l'évaluation psycholinguistique 236
F. Annexe 6 : WordNet 237
G. Annexe 7 : listes de cliques 239
Index 246
Bibliographie 248

Table des figures

Figure 3.1 : Arbre hiérarchique de décision *47*
Figure 3.2 : Résultats obtenus sur les dix mots testés *53*
Figure 3.3 : Exemple d'entrée du dictionnaire CTRW *57*
Figure 3.4 : Extrait de résultat du calcul de similarité *64*
Figure 3.5 : Traitement d'un groupe de noms *65*
Figure 3.6 : Exemple d'entrée de la ressource lexicale WordNet *78*
Figure 4.1 : Structure de la définition de monter *dans le TLFi* *88*
Figure 4.2 : Structure de la définition de compter *dans le TLFi* *90*
Figure 4.3 : Structure de la définition de jouer *dans le TLFi* *92*
Figure 4.4 : Espace sémantique du verbe monter *98*
Figure 4.5 : Espace sémantique du verbe compter *101*
Figure 4.6 : Espace sémantique du verbe jouer *103*
Figure 4.7 : Fonctions potentielles des synonymes manigancer, augmenter *et* s'élever *sur l'espace sémantique du verbe* monter *105*
Figure 5.1 : Représentation d'une fonction potentielle sur un espace sémantique bidimensionnel *107*
Figure 5.2 : Les différents cas de figure interprétatifs (espace unidimensionnel) *108*
Figure 6.1 : Interface pour les requêtes Frantext *123*
Figure 6.2 : Structure algorithmique de la fonction requeteFrantext *124*
Figure 6.3 : Fonction lancerRequeteFrantext *124*
Figure 6.4 : Récapitulatif des fréquences dans Frantext *126*
Figure 6.5 : Fonction potentielle du co-texte escalier *126*
Figure 6.6 : Fonction potentielle du co-texte diamant *130*
Figure 6.7 : Fonction potentielle du co-texte projet *133*
Figure 7.1 : Représentation de la construction ditransitive *145*
Figure 7.2 : Représentation de la construction ditransitive couplée avec le verbe hand *146*
Figure 7.3 : Représentation de la construction de mouvement induit *148*
Figure 8.1 : Espace sémantique du verbe compter *(rappel)* *157*
Figure 8.2 : Liste de synonymes de compter *par ordre de fréquence d'emploi avec la construction* V sur SN *158*
Figure 8.3 : Liste de synonymes de compter *par ordre de degré d'affinité avec la construction* V sur SN *158*
Figure 8.4 : Liste des cliques contenant compter *par ordre de degré d'affinité avec la construction* V sur SN *158*
Figure 8.5 : Fonction potentielle de la construction V sur SN *sur l'espace sémantique de* compter *159*
Figure 8.6 : Fonction potentielle de la construction V SN *sur l'espace sémantique de* compter *159*
Figure 8.7 : Interface pour les requêtes Syntex *161*

Figure 8.8 : Fonction récursive pour la construction de requête Perl *163*
Figure 8.9 : Fonction potentielle de V sur SN *165*
Figure 8.10 : Fonction potentielle de V sur SN *166*
Figure 8.11 : Fonctions potentielles de V de SN *et* V à SN *168*
Figure 8.12 : Exemple de taux d'adéquation entre un synonyme et une construction égal à zéro *174*
Figure 8.13 : Résultats de l'évaluation psycholinguistique *177*
Figure 8.14 : Espace sémantique du verbe jouer *(rappel)* *181*
Figure 8.15 : Fonction potentielle du co-texte fille *181*
Figure 8.16 : Combinaison des fonctions potentielles du co-texte fille *et de la construction* V *SN* *182*
Figure 8.17 : Combinaison des fonctions potentielles du co-texte fille *et de la construction* V avec SN *182*
Figure 9.1 : Extrait du système de requête automatique en php *193*
Figure 9.2 : Tableau récapitulatif des IM calculés *193*
Figure 9.3 : Visualisation des résultats *197*
Figure 9.4 : Distribution des mots compléments d'objet de descendre auxquels on a ajouté Seine *et* Mont-blanc *205*
Figure 9.5 : Extrait des notes d'évaluation *208*
Figure 9.6 : Evaluation des classes obtenues pour 4 contextes, 60 cooccurrences, 8 juges *209*
Figure 9.7 : Différentes classes de Wimbledon *en fonction du contexte* *210*
Figure 10.1 : Fonctions potentielles obtenues à partir du corpus Frantext *216*
Figure 10.2 : Fonctions potentielles obtenues à partir du corpus LM3 ..... *216*
Figure 10.3 : Processus de sélection de paraphrases d'une construction verbale *221*

1. Introduction

1.1. Objectif de la thèse

Les défis de la société de l'information

La modélisation des processus linguistiques est un enjeu déterminant et d'actualité. La société dans laquelle nous vivons est souvent décrite comme une société de l'information. Nous devrions plutôt parler de société de sur-information, c'est-à-dire une société dans laquelle la recherche d'information ne correspond pas à la question « Y a-t-il une réponse à ma question ? » mais plutôt « Où puis-je trouver les réponses à ma question ? ». Une solution pour rendre cette information accessible à tous, c'est-à-dire plus démocratique, est de développer un ensemble d'outils capables de rechercher, de trier, d'indexer, de résumer, de traduire cette information. Actuellement, la plupart de ces tâches sont effectuées à l'aide d'outils statistiques classiques, du type *data mining*, sans aucune ressource linguistique. Or, les limites de ces outils sont très vite atteintes lorsque l'on a affaire à de la compréhension de texte comme c'est particulièrement le cas pour la traduction, la synthèse d'information et plus généralement pour l'ensemble des tâches de manipulation de texte. C'est dans ce cadre que la linguistique a un rôle déterminant à jouer. Il ne s'agit plus de savoir si un mot, un *pattern* ou un ensemble de *patterns* est présent dans le texte, mais bien de comprendre le sens d'un texte.

Qu'est-ce que comprendre un texte ?

Pour les humains comme pour les ordinateurs, il y a plusieurs niveaux de compréhension. Pour le lecteur humain, ce n'est pas pareil de survoler rapidement un article pour savoir simplement de quoi il parle, ou encore pour y rechercher une information précise dont on a besoin, et d'étudier en détail un livre de cours ou un essai.

C'est encore différent de se plonger dans un roman ou de savourer de la poésie… Pour un système, on a aussi des différences pour les tâches correspondantes : indexation et extraction d'information réclament une « compréhension » plus légère que la construction d'une base de connaissances à partir d'un ouvrage technique, par exemple. Quant à l'équivalent de la lecture d'un roman ou d'une poésie, c'est pour l'instant hors de portée…

Le rôle du contexte dans la compréhension

Pour l'humain, la compréhension d'un énoncé utilise massivement le contexte de cet énoncé. Ce contexte consiste en plusieurs éléments : les connaissances générales qu'il a sur le monde, ce qui entoure le texte lui-même, et la partie du texte qui précède l'énoncé en question.

Par exemple, un énoncé comme *Les bleus ont plutôt mieux réussi leurs paniers à trois points que les jours précédents* ne posera pas de problème de compréhension parce que le lecteur le lira dans un journal, dans la page sports, peut-être même sous la rubrique basket, que le début du texte parlera d'un match ayant eu lieu la veille dans laquelle jouait l'équipe de France, qu'il saura que la couleur des maillots de cette équipe est le bleu et qu'il aura quelques notions des règles du jeu du basket…

Pour l'ordinateur, le rôle du contexte sera bien entendu tout aussi important. Par exemple, pour une tâche d'extraction d'information, on utilisera le maximum de connaissances sur le domaine précis sur lequel on travaille ; comme on le sait bien, le travail sur un corpus homogène, dans un domaine de spécialité donné, est plus facile : c'est, d'une certaine manière, parce qu'on maîtrise mieux une partie du contexte des énoncés que l'on a à analyser. Pour reprendre notre exemple, le mot *panier* aura presque systématiquement le sens qu'il a au basket dans les articles de sport. Faire l'impasse sur la polysémie de ce mot pour ce genre de textes ne produira que très peu de bruit.

Mais le contexte n'est pas tout

S'il ne faut donc pas négliger le contexte dans les tâches de compréhension automatique, il ne faut pas non plus faire l'erreur inverse qui consisterait à penser que les connaissances données par le contexte suffisent pour l'essentiel, et que l'on peut éviter d'utiliser des connaissances linguistiques plus approfondies dans ce genre de tâches. En reprenant notre exemple d'article sur le basket, on peut montrer qu'un certain nombre d'ambiguïtés lexicales demeurent, même si l'on sait que l'on a affaire à un texte sur ce sujet. Prenons par exemple le verbe *tirer* : le sens de 'shooter', auquel on pense tout de suite, est loin d'être le seul qu'on peut trouver dans ce contexte. Exemples : *L'entraîneur a su tirer le maximum de son équipe* ; *alors que le match tirait à sa fin, …* ; c*e joueur a tiré vers le haut toute l'équipe* ; etc. De même, un nom comme *faute* peut avoir plusieurs sens, en plus de la faute de jeu (*Ce joueur va sortir pour cinq fautes*) : *Cet échec est entièrement de la faute de l'entraîneur* ; *faute de motivation, les joueurs n'ont jamais trouvé le bon rythme* ; etc.

Pour traiter correctement ces emplois, il faut analyser l'énoncé lui-même, et cela réclame donc des connaissances sur la langue, que partagent tous les locuteurs, et qui sont valides dans tous les contextes : c'est ce que nous appelons *la compétence linguistique* d'un locuteur.

Augmenter la compétence linguistique des systèmes de traitement : un enjeu important pour la compréhension automatique

Pour améliorer les systèmes de compréhension, et donc répondre aux défis de la société de l'information que nous avons évoqués au début de cette introduction, on ne peut se contenter de mieux prendre en compte le contexte et les connaissances du domaine d'étude : il faut aussi augmenter la compétence linguistique des systèmes. C'est indispensable pour progresser vraiment.

Beaucoup de progrès ont été réalisés ces dernières années, notamment dans le domaine de la morphologie et de la syntaxe. On dispose ainsi aujourd'hui d'outils assez performants dans ces domaines : les étiqueteurs morphosyntaxiques deviennent de plus

en plus fiables (Cordial, Tree Tagger, Brill, etc.), et depuis peu des analyseurs syntaxiques tels que Syntex, Xip ou encore le Lasaf donnent des résultats très satisfaisants.

Mais la sémantique n'a pas progressé au même rythme. Que ce soit pour la traduction de textes, la synthèse de texte, ou encore l'extraction d'information, il n'existe pas de système automatisé satisfaisant à l'heure actuelle.

Objectif de la thèse : contribuer à augmenter la compétence linguistique des systèmes dans le champ de la sémantique lexicale

Le retard pris en sémantique lexicale peut s'expliquer notamment par l'omniprésence de la polysémie. En effet, la plupart des mots possèdent plusieurs sens et l'ambiguïté d'un mot ne peut être réduite que par son contexte d'emploi. Notre travail a consisté à développer un outil de désambiguïsation automatique des mots. Il contribue ainsi à l'augmentation des compétences linguistiques des systèmes de traitement en sémantique lexicale.

Il s'agit avant tout d'un travail de linguistique informatique et non d'informatique linguistique, c'est-à-dire un travail qui consiste à utiliser l'informatique « comme outil de validation d'hypothèses théoriques sur le fonctionnement de la langue »[1] (Fuchs *et al.*, 1993), et non à appliquer des outils informatiques à des données linguistiques. Il s'agit d'un travail sur la langue qui ne vise pas directement une application particulière. Cependant, notre souci constant a été de développer un modèle automatique qui puisse être facilement utilisé dans des applications. Notamment, il y a

1 Définition complète proposée par Fuchs *et al.* (1993 : 21) : « La linguistique informatique (en anglais computational linguistics), quant à elle, […] recourt à l'ordinateur comme outil de validation d'hypothèses théoriques sur le fonctionnement de la langue : elle a vocation à se servir de techniques informatiques pour expérimenter sur des données linguistiques, tester des systèmes de règles, étudier la pertinence linguistique de certaines modélisations, voire simuler certains comportements langagiers. [Relèvent de la] « linguistique informatique » tous les travaux en TAL qui, d'une manière ou d'une autre, s'appuient sur des éléments d'analyse linguistique. »

eu un effort important pour que notre modèle repose sur des méthodes de calcul réalistes et nécessitant des quantités de ressources raisonnables.

Notre objectif à long terme est le calcul du sens d'un énoncé. Mais l'omniprésence de la polysémie implique de devoir aussi calculer le sens des unités de l'énoncé. Calculer le sens d'un verbe est une clé importante du problème puisque maîtriser la sémantique de la construction verbale est une étape cruciale pour la compréhension de l'énoncé. Or nous allons voir que les verbes ont été peu ou mal traités en désambiguïsation automatique (chapitre 3). Nous proposons un modèle de désambiguïsation des verbes qui tente de mieux rendre compte de la polysémie verbale et de ses singularités, notamment en donnant une place nouvelle à la construction syntaxique, fondée sur les grammaires constructionnelles. Toujours dans l'objectif du calcul du sens d'un énoncé, nous proposons une méthode de calcul du sens d'une construction verbale.

1.2. Problématique

La désambiguïsation automatique fait partie des domaines du Traitement Automatique de la Langue (TAL) les plus productifs[2]. Deux raisons expliquent cette productivité : d'une part l'intérêt considérable de la désambiguïsation automatique pour un nombre important d'applications, d'autre part l'absence de consensus sur la manière d'appréhender cette tâche. Nous verrons, notamment dans le chapitre 3, que les approches diffèrent sur de nombreux points : ressource à utiliser, information à extraire de ces ressources, résultats à produire, etc.

Notre objectif est de présenter un modèle de désambiguïsation que nous pensons d'un genre nouveau. Pour cela, nous proposons de poser à plat quelques

[2] Pour donner une idée de la productivité de cette discipline, signalons que le moteur de recherche *Scholar Google* renvoie à 244 articles contenant l'expression « word sense disambiguation » uniquement pour l'année 2004.

questions relatives à la tâche de désambiguïsation. Qu'est ce que le sens d'un mot ? Qu'est ce que le sens d'un énoncé ? Comment les représenter ?

En présentant l'objectif général de cette thèse, nous avons insisté sur le fait que nous nous cantonnons à ce que nous avons appelé une tâche de « compétence linguistique ». Nous ne répondrons à ces questions qu'en terme de « compétence linguistique ». En effet, notre objectif est de construire un outil capable de donner un sens à un énoncé uniquement à l'aide des données linguistiques contenues dans cet énoncé, que nous nommerons dorénavant le co-texte. Ce qui implique déjà des processus cognitifs complexes.

Intéressons-nous un instant à l'aspect cognitif de la tâche que nous nous sommes fixé. Il ne fait aucun doute que le processus de compréhension de texte est un processus très complexe du point de vue cognitif. Comprendre un énoncé implique la considération du « contexte d'énonciation », c'est-à-dire notamment des caractéristiques du locuteur, du lieu d'énonciation, ou encore de la déictique. Prenons maintenant les deux énoncés suivants :

(1) Monter les escaliers

(2) Monter un projet

Ces deux énoncés ne sont pas inclus dans un contexte d'énonciation et pourtant, chacun s'accordera sur le fait que le verbe *monter* n'a pas le même sens dans l'énoncé (1) (*gravir, grimper*) que dans l'énoncé (2) (*élaborer, réaliser*). Il y a donc bien un processus cognitif sous-jacent nous permettant de distinguer le sens d'un mot uniquement en fonction de l'énoncé dans lequel il se trouve. Notre système n'a cependant aucune ambition de modélisation ni neurophysiologique ni psycholinguistique, c'est-à-dire que si les processus que nous cherchons à modéliser sont bien cognitifs, nous ne cherchons pas à fonder notre modèle sur les connaissances que l'on a du fonctionnement cognitif de notre cerveau. Notre objectif est simplement

d'implémenter un modèle mathématique, fondé sur un modèle linguistique, capable d'accomplir la tâche de calcul du sens d'un énoncé.

Pour cela, nous proposons de nous appuyer sur le modèle de la construction dynamique du sens développé par Victorri et Fuchs (1996). Nous développerons ce modèle dans le chapitre 4. L'intérêt de ce modèle est qu'il permet de répondre aux questions que nous avons posées en début de paragraphe en respectant notre cadre d'étude qui est la « compétence linguistique ». Nous proposerons ainsi de définir le sens en passant par la polysémie (cf. chap. 2) et plus précisément de représenter, d'une part les sens d'un mot à l'aide d'un espace sémantique, et d'autre part les sens d'un énoncé à l'aide de paraphrases.

Etudions maintenant les types de résultats que nous attendons et les critères linguistiques que cela implique.

Nous venons de proposer de décrire le sens d'un énoncé par un ensemble d'énoncés ayant le même sens ou un sens très proche. Cela correspond à une représentation assez réduite du sens d'un énoncé mais qui a le mérite d'être évaluable par des critères linguistiques. Calculer les paraphrases d'un énoncé entier est un objectif à long terme. Nous proposons ici de décrire une étape importante de cet objectif : calculer le sens de constructions verbales en construisant automatiquement des paraphrases de ces constructions verbales.

Par exemple, nous voulons que notre outil soit capable de dire que le sens de la construction verbale *jouer de la guitare* est très proche du sens de la construction verbale *pratiquer la guitare*. Retrouver cette proximité de sens implique de pouvoir traiter certains des phénomènes linguistiques les plus fondamentaux. Le premier d'entre eux est la polysémie, c'est-à-dire le fait que la grande majorité des mots de la langue peuvent avoir plusieurs sens (nous développerons ce que nous entendons précisément par *polysémie* dans le chapitre 2). C'est bien ce phénomène qui rend le calcul du sens d'un énoncé si complexe. Reprenons notre exemple *jouer de la guitare*. Le verbe *jouer* est un verbe particulièrement polysémique, le TLFi en propose une définition de 16

pages décrivant 116 acceptions. Il peut notamment prendre le sens de *parier* (*jouer sur un cheval*), de *s'amuser* (*jouer dans la cour*) ou encore d'*interpréter* (*jouer Andromaque*). La première difficulté est de représenter cette pluralité de sens tout en respectant l'aspect perméable des frontières entre sens : *jouer Andromaque* et *jouer l'idiot* illustrent deux sens différents de *jouer*, respectivement *interpréter un rôle* et *imiter une attitude*, mais la frontière sémantique entre imitation et interprétation est difficile à établir. Il nous faut donc un modèle de représentation du sens capable de rendre compte de la polysémie et de cette continuité entre les sens.

En considérant que ce modèle existe, il faut, pour chaque unité de la construction verbale étudiée, être capable de retrouver parmi ses sens possibles le ou les sens appropriés. Une première idée serait de remplacer chaque mot par un synonyme approprié. Par exemple, un bon synonyme de *jouer* dans la construction verbale *jouer de la guitare*, serait *pratiquer*. Cette étape pourrait nous satisfaire si l'on ne s'intéressait qu'au verbe de départ, ici *jouer*. En effet, *pratiquer*, en tant que synonyme de *jouer*, ne peut avoir que le sens de « *pratiquer une activité* ». Mais indépendamment de *jouer*, *pratiquer* peut aussi prendre le sens de « *fréquenter/emprunter habituellement* » (*il pratique un chemin/une route*, cf. TLFi) ou encore « *appliquer*, *mettre en action une méthode, une théorie* » (*Les auteurs qui pratiquent la méthode objective*, cf. TLFi). Et même si l'on reste dans les sens de *pratiquer* en tant que synonyme de *jouer*, il est possible de distinguer deux sens de *pratiquer* qui sont *pratiquer un instrument de musique* et *pratiquer un sport*. Autrement dit, proposer un seul verbe comme synonyme de *jouer* dans *jouer de la guitare* semble insuffisant. Cela voudrait dire que l'ensemble des sens de *pratiquer* convient pour remplacer *jouer* dans *jouer de la guitare*. Or, **appliquer la guitare* ne se dit pas et *emprunter la guitare* a un tout autre sens que *jouer de la guitare*. Même si l'on reste dans les sens communs à *pratiquer* et *jouer*, il faut encore être capable de déterminer si l'on parle de *pratiquer* au sens de *s'entraîner à un sport*, ou au sens de *manier un instrument*, *utiliser un instrument*. L'idée, pour résoudre

ce problème, serait de proposer un ensemble de paraphrases, plutôt qu'une seule, classées par ordre de pertinence.

Une autre difficulté du paraphrasage concerne la construction syntaxique. *Pratiquer* est un bon synonyme de *jouer* dans *jouer de la guitare* à condition que l'on dise *pratiquer la guitare* et non **pratiquer de la guitare*. Cela nous apprend deux choses : premièrement qu'il ne suffit pas de trouver un bon synonyme, mais il faut aussi trouver la construction syntaxique qui convient à ce synonyme. Cela semble deuxièmement nous montrer que la construction syntaxique n'est pas déterminante dans la désambiguïsation : il est difficile de trouver un sens commun entre *jouer de son charme*, *jouer de la guitare* et *jouer des coudes*, qui ont pourtant la même construction. A l'inverse, l'intuition nous fait dire que *jouer de quelque chose* n'a pas le même sens que *jouer quelque chose*. Certains verbes comme *monter* peuvent avoir des sens très différents pour une même construction syntaxique : *monter les escaliers (gravir, grimper) / monter un projet (élaborer, réaliser)* alors que d'autres ont des constructions extrêmement tranchées comme le verbe *compter* : *compter sur quelque chose* (*s'appuyer, tabler*) n'a pas du tout le même sens que *compter quelque chose* (*énumérer, dénombrer*). Autrement dit, s'il n'y a clairement pas de bi-univocité entre constructions syntaxiques et sens, les critères syntaxiques doivent sans aucun doute être pris en compte dans notre modèle.

Enfin, un dernier point à traiter est la généralisation de nos paraphrases : le fait que *jouer de la guitare* ait le même sens que *pratiquer la guitare* n'est pas valable seulement pour *guitare* mais pour tous les instruments de musique. Autrement dit, il serait dommage de chercher des paraphrases pour la construction verbale *jouer de la guitare*, indépendamment de *jouer du piano, du violon, etc.* puisque la relation de paraphrase entre *jouer de quelque chose* et *pratiquer quelque chose* est valable pour tous les instruments de musique. Par conséquent, il semble important, avant de rechercher des paraphrases, de connaître les caractéristiques sémantiques des unités de

la construction verbale, toute la difficulté étant de trouver une méthode capable de calculer la ou les classe(s) sémantique(s) d'une unité.

Pour résumer, l'objectif est de proposer un système capable de calculer automatiquement le sens d'une construction verbale, c'est-à-dire de donner un sens à chaque unité lexicale de cette construction verbale (nous verrons comment dans les chapitres 4 et 5), et de donner un sens à la construction verbale elle-même en proposant une série de paraphrases. A partir des différents points que nous venons de soulever, notre calcul du sens devra tenir compte de l'influence des unités lexicales de la construction verbale, de l'influence de la construction syntaxique ainsi que de l'influence des caractéristiques sémantiques des unités de la construction verbale.

1.3. Plan de la thèse

Puisque la polysémie est la source de tous nos « mots », nous commencerons par présenter ce phénomène (Chapitre 2). Nous montrerons notamment qu'il n'existe pas une mais des polysémies et nous préciserons le type de polysémie sur lequel nous avons travaillé. Le chapitre 3 propose un état de l'art un peu particulier sur la désambiguïsation lexicale automatique. Il se décompose en trois parties : la première consiste à décrire les différentes approches par l'intermédiaire de travaux représentatifs. La deuxième partie décrit quelques problèmes récurrents concernant l'évaluation des modèles de désambiguïsation. La troisième partie présente les sources de connaissances que nous utilisons dans notre modèle de calcul du sens. Nous replaçons ensuite notre modèle dans ce cadre et présentons son fondement théorique ainsi que les développements logiciels disponibles, notamment pour la représentation d'un espace sémantique (chapitre 4). Les chapitres suivants correspondent aux développements effectués durant cette thèse. Nous commencerons par présenter la méthode, notamment les corpus utilisés et les méthodes de calculs utilisés (chapitre 5). Nous passerons ensuite au modèle de désambiguïsation à proprement parler, avec une première partie sur l'influence du co-texte lexical (chapitre 6) et une seconde sur l'influence du co-texte

syntaxique (chapitre 8). Ce chapitre sera précédé d'une présentation détaillée des grammaires constructionnelles (chapitre 7). Cette grammaire correspond en effet à la base théorique de notre approche des constructions syntaxiques. Nous présenterons dans le chapitre 9 deux pistes de recherche relatives à certains problèmes rencontrés par notre modèle de désambiguïsation. Enfin nous conclurons en brossant à grands traits les perspectives et applications qui nous semblent réalisables à partir de notre modèle (chapitre 10).

2. La polysémie

Dans ce chapitre, nous allons d'abord chercher à mieux cerner ce que nous entendons par polysémie en l'opposant à d'autres phénomènes linguistiques, que nous excluons de notre champ d'étude[3]. Nous présenterons ensuite deux conceptions diamétralement opposées de la polysémie afin de délimiter dans la mesure du possible le panel dans lequel se trouvent toutes les études de ce phénomène. La première conception consiste à considérer la polysémie comme un artefact, une pure création de linguistes afin d'expliquer les différents sens que les mots peuvent prendre : la polysémie ne correspondrait alors pas à un phénomène réel faisant partie des langues naturelles. La deuxième conception consiste au contraire à traiter la polysémie comme un « passage obligé » dans l'élaboration du sens de n'importe quelle unité de la langue. Enfin, nous proposerons de pointer sur quelques particularités de la polysémie verbale, qui est au centre de cette thèse.

2.1. Polysémie et homonymie : continuités et ruptures de sens

Partons de la définition de la polysémie proposée par Kleiber (1999 : 55) :

(i) une pluralité de sens liée à une seule forme

(ii) des sens qui ne paraissent pas totalement disjoints, mais se trouvent unis par tel ou tel rapport.

Il oppose cette définition à celle de l'homonymie où seule (i) est valable. Ainsi dira-t-on que *plateau* est polysémique parce qu'il prend des sens différents dans *un*

[3] Les trois premières sections de ce chapitre reprennent dans ses grandes lignes la première partie de l'article Jacquet *et al.* 2005.

plateau à fromages et *un plateau de théâtre*, et que ces sens sont unis par l'évocation commune d'une forme horizontale sur laquelle peuvent être disposées un certain nombre de choses. En revanche, on parlera d'homonymie dans le cas de *canon* parce que ses deux sens dans *un tir au canon* et *le droit canon* sont totalement disjoints. Alors que *plateau* apparaît bien comme une unité lexicale unique malgré sa pluralité de sens, on aura tendance à considérer dans le cas de *canon* que l'on a affaire à deux unités distinctes, $canon_1$ et $canon_2$, qui partagent une même forme. Il faut noter cependant que la frontière entre polysémie et homonymie n'est pas très nette, comme en témoignent le caractère assez vague de la formulation de la deuxième caractéristique (ii) et les divergences des dictionnaires sur le nombre d'entrées consacrées à telle ou telle unité. Ainsi, faut-il traiter *table* comme une seule unité polysémique, ou doit-on distinguer un $table_2$ à l'œuvre dans *table de multiplication* du $table_1$ qu'on trouve dans *table de cuisine* ?

Plutôt que de chercher à trancher de manière inévitablement arbitraire, il vaut sans doute mieux accepter l'existence d'un continuum et adopter une définition de la polysémie qui en tienne compte (Victorri et Fuchs, 1996). Partant du fait que le sens d'une unité lexicale dans une occurrence donnée dépend en partie de ce qu'apporte cette unité elle-même de constant, quel que soit le contexte, et en partie de ce qui est fonction du contexte, on peut classer les expressions suivant l'importance relative de ces deux facteurs. A un extrême, le contexte ne joue aucun rôle : l'expression est monosémique ; son sens est le même dans tous les énoncés, ce sens étant donc entièrement défini par l'apport propre de l'unité (exemples : *tournevis*, *hectolitre*, …). A l'autre extrême on trouve les homonymes « purs », dont l'apport constant, commun à tous les emplois, est effectivement nul, puisque le sens peut changer radicalement suivant les énoncés (exemples : *avocat*, *sol*, …). Entre ces deux extrêmes, se trouve le cas général de la polysémie, avec des cas qui tendent vers la monosémie, quand le contexte ne joue qu'un rôle minime (tous les sens recensables sont très proches les uns des autres), et d'autres vers l'homonymie, quand l'apport propre constant est très faible.

Prenons quelques exemples pour illustrer ce dernier point. Soit le mot *bureau*. Il possède quatre sens principaux : un meuble (ex. : *s'asseoir à son bureau*), une pièce (ex. : *ouvrir la fenêtre de son bureau*), un établissement (ex. : *le bureau de poste*, *le bureau de tabac*, etc.), une institution (ex. : *le bureau de l'Assemblée*, *le bureau de l'association*, etc.). Ces différents sens sont indéniablement reliés (comme on le verra ci-dessous, ils sont dans des rapports de *métonymie*), ce qui signifie que l'on a bien affaire à de la polysémie et non de l'homonymie. Cependant, quand on essaie de déterminer l'apport propre du mot *bureau* qui est commun à tous ses emplois, on s'aperçoit qu'il est très ténu : une vague notion d'activité d'écriture. Autrement dit, on se trouve plus près du pôle homonymique dans le continuum qui va de la monosémie à l'homonymie.

A l'inverse, prenons le mot *livre*. Il a aussi plusieurs sens : il peut désigner notamment un objet physique (*un petit livre*, *de couverture rouge, posé sur l'étagère*), une production intellectuelle (*un livre très drôle*, *mais très mal écrit*), un produit commercial (*un livre trop cher et introuvable, bien que paru récemment*). Mais l'apport propre de ce mot, commun à ces différents empois, est ici beaucoup plus important. On peut facilement utiliser une notion « générique » de livre, en tant qu'entité abstraite, possédant un contenu intellectuel, produit à de multiples exemplaires pouvant être vendus. Cette notion générique peut ensuite être spécifiée pour aboutir à un sens précis dans un contexte particulier. On a donc affaire dans ce cas à une polysémie beaucoup plus proche de la monosémie que de l'homonymie.

Il faut noter que la situation est généralement plus complexe. Prenons l'exemple du mot *pied*. On peut trouver dans un grand nombre de ses emplois un apport de ce mot que l'on peut formuler de la manière suivante : il désigne la partie la plus basse d'un objet physique, en contact avec une surface horizontale[4]. Cela s'applique bien sûr à

4. Il ne faut pas prendre cet « apport » du mot pour une sorte de définition. En l'occurrence, notre formulation n'est pas une définition de *pied* parce que bien des parties basses en contact avec le sol ne peuvent pas être désignées par *pied*, à commencer par les pattes des animaux.

l'extrémité du corps humain que l'on appelle ainsi, mais aussi au pied de la table, de l'arbre, du verre, de la falaise, etc. Et cela semble d'autant plus important de caractériser cet apport propre qu'il est très productif : il existe des centaines de noms N pour lesquels le sens de l'expression *pied de N* peut être calculé en utilisant cette formulation. Mais il y a aussi d'autres sens du mot *pied* pour lesquels cette notion de partie basse n'est pas pertinente, comme le sens d'unité de mesure (*haut de six pieds*) ou celle d'unité rythmique (*un vers de six pieds*)[5]. On a donc intérêt, pour des cas comme celui-ci (et ils sont nombreux), à concevoir un modèle à plusieurs niveaux : un niveau, proche de l'homonymie, qui opère une première séparation entre des sens trop éloignés pour que la caractérisation d'un apport commun soit utilisable, et un deuxième niveau, où l'on peut, pour chaque sous-ensemble de sens, rendre opérationnelles des formulations de l'apport propre qui facilitent le calcul du sens exact de l'unité dans un énoncé donné.

2.2. Phénomènes de parole et degrés de lexicalisation

La polysémie est un phénomène « vivant », au sens où les mots sont en permanence susceptibles d'acquérir de nouveaux sens (et aussi d'en perdre d'autres) : c'est l'un des moteurs les plus importants de l'évolution du lexique d'une langue. Deux grands procédés sont principalement à l'origine de l'acquisition de ces nouveaux sens : la *métaphore* et la *métonymie*. La métaphore consiste à utiliser un mot qui désigne habituellement une entité ou un événement d'un certain domaine pour évoquer une entité ou un événement qui joue un rôle analogue dans un autre domaine. Par exemple, c'est par métaphore que l'on parle de *virus informatique* : le mot *virus*, qui vient du domaine de la biologie, est utilisé pour parler de programmes informatiques dont le comportement rappelle celui des virus biologiques. Quant à la métonymie, c'est le procédé qui consiste à évoquer une entité (ou un événement) par le mot qui désigne une autre entité (ou événement), liée à la première par un rapport fonctionnel ou structurel.

5. Sans parler de locutions figées, comme *prendre son pied*, *au pied de la lettre* ou *faire le pied de grue*, qui peuvent être traitées comme des unités à part entière.

Ainsi c'est par métonymie que *le premier violon* désigne un violoniste, ou que l'on dit *faire rire la salle* alors que ce sont les occupants de la salle qui rient. Pour reprendre les exemples examinés ci-dessus, c'est par métonymies successives que le mot *bureau* a acquis ses différents sens au cours de l'évolution du français, passant du meuble à la pièce contenant ce meuble, puis au lieu constitué de telles pièces, puis aux groupes travaillant dans de tels lieux[6], alors que c'est par métaphore que le pied, partie de corps humain, a pu aussi désigner la partie correspondante d'une table ou d'un arbre[7].

La métaphore permet de produire des sens nouveaux de manière quasiment illimitée au cours du discours : on parle de « métaphore vive » (Ricoeur, 1975). La métaphore vive est une création éphémère de la parole. C'est « un rapprochement soudain entre des choses qui semblaient éloignées » (Ricoeur, 1975 : 49) :

Son bureau est un hall de gare

La structure du chromosome est tout à la fois code législatif et pouvoir exécutif

Il en est de même pour la métonymie. Ainsi, on caractérise comme des *métonymies vives* les emplois suivants (Nunberg, 1978 ; Fauconnier, 1984) :

L'omelette au jambon est parti sans payer !

(énoncé par un serveur de restaurant s'adressant au cuisinier)

L'appendicite du troisième a encore fait de la fièvre cette nuit.

(une infirmière à un médecin à l'hôpital).

Mais un certain nombre de ces emplois « passent dans la langue », progressivement, au sens où ils deviennent d'emploi banal, au point d'être répertoriés dans le dictionnaire. On parle alors de métaphores ou de métonymies *lexicalisées* ou

6. Dans son premier sens, perdu aujourd'hui, il désignait une étoffe (celle de la bure des moines) que l'on plaçait sur les tables sur lesquelles on lisait et écrivait : le sens de meuble est donc déjà le résultat d'une première métonymie.

7. En revanche, c'est par métonymie qu'il a acquis son sens d'unité de mesure (longueur évaluable avec son pied).

conventionnelles. Ainsi, dans *J'ai une montagne de choses à faire*, le sens métaphorique de *montagne* n'est plus ressenti comme une figure de rhétorique, mais comme un sens de plein droit du mot *montagne*. De même pour l'emploi métonymique de *bouteille* dans *boire une bonne bouteille.* Ce processus de lexicalisation peut aller jusqu'à la perte totale, dans la conscience des locuteurs, de l'existence même d'un trope à l'origine du sens dérivé. On parle alors de métaphore ou de métonymie « morte ». Ainsi *voler* au sens de dérober et *voler* au sens de se déplacer dans les airs sont considérés aujourd'hui comme des homonymes, le rapport de sens entre eux s'étant complètement perdu, même si, en fait, le premier dérive du second par un processus métaphorique (on disait en français classique : *Le faucon vole sa proie*).

Là encore, il existe bien des cas intermédiaires, entre métaphore (ou métonymie) vive et lexicalisée, ou lexicalisée et morte, qui sont le reflet en synchronie de l'état plus ou moins avancé des processus diachroniques qui tendent à banaliser de plus en plus un certain nombre d'inventions discursives des locuteurs, jusqu'à ce que leur origine devienne parfaitement opaque.

Peut-on calculer des sens métaphoriques ou métonymiques ? Le problème se pose de manière particulièrement aiguë pour les sens non lexicalisés, qui, par définition, ne sont pas recensés dans les dictionnaires. En effet, la tâche est dans ce cas beaucoup plus ardue : le modèle de calcul du sens ne consiste pas simplement à sélectionner un sens pertinent parmi un ensemble de sens déjà répertoriés ; il faut « découvrir » le nouveau sens produit par la métaphore ou la métonymie vive. Ainsi Duvigneau (2002 ; 2003) a trouvé dans un texte de physique de Poincaré l'exemple suivant : *Faudra-t-il chercher à raccommoder les principes ébréchés (…)* ? Les sens de *raccommoder* et de *ébréchés* dans ce contexte n'ont aucune chance de se trouver dans le lexique. Duvigneau analyse ce type de métaphores verbales comme un processus de co-hyponymie (le terme métaphorique *raccommoder* et les termes plus conventionnels *réviser*, *remanier* étant des hyponymes d'un même hyperonyme *réparer*). On a donc dans ce cas une méthode

de calcul du sens métaphorique, qui semble implémentable si l'on dispose d'une ressource lexicale fournissant les relations d'hyperonymie[8].

Malheureusement, toutes les métonymies et les métaphores vives ne sont pas susceptibles d'un tel traitement. Notamment, les exemples que nous avons donné au début de cette section, qu'il s'agisse du *code législatif* de Ricœur ou de *l'omelette au jambon* de Fauconnier, montrent que l'on a souvent besoin d'un contexte très large (incluant la situation d'énonciation) et de connaissances précises sur le monde pour pouvoir interpréter correctement ces expressions. Il faut donc admettre que l'on se trouve là au-delà de ce dont on est capable en traitement automatique, aujourd'hui et vraisemblablement pour encore de longues années…

2.3. Métonymies intégrées et coercition de type

Un certain nombre de phénomènes de changement de sens sont cependant suffisamment systématiques pour pouvoir être traités par des règles générales. C'est le cas notamment de ce que Kleiber (1994 ; 1999) a appelé les *métonymies intégrées*. Prenons les exemples suivants :

Je suis garé sur la place du marché

George Sand est sur l'étagère de gauche

Paris a massivement voté « oui » au referendum

Il est clair que ce n'est pas « moi » qui suis garé sur la place, mais ma voiture. De même ce n'est pas George Sand mais un exemplaire d'un livre dont elle est l'auteur qui a été rangé sur l'étagère, et ce n'est pas Paris qui a participé au référendum mais bien ses habitants. Dans le premier exemple, on peut d'ailleurs remplacer *je* par *mon frère*, *l'un des suspects* ou *l'épicier du coin*, ce sera toujours d'un véhicule qu'il s'agira,

8. Cf. le travail de Gaume (Gaume *et al.*, 2002 ; Gaume 2003), qui a conçu une mesure sur des graphes lexicaux, la *proxémie*, qui permet, entre autres, d'obtenir automatiquement ces co-hyponymes.

ce qui montre que ces phénomènes sont vraiment systématiques : c'est toute la classe des groupes nominaux désignant des humains qui possède la capacité de désigner un véhicule dans la position sujet de *être garé* et de bien d'autres prédicats (*rouler*, *déraper*, *entrer en collision*, etc.). Il serait bien sûr complètement inefficace de vouloir conserver dans le lexique cette information pour toutes les unités lexicales concernées. Il faut au contraire entrer ces règles générales en tant que telles avec des mécanismes de déclenchement de type « résolution de conflits »[9].

Pour traiter ces phénomènes, plusieurs auteurs ont cherché à définir des mécanismes généraux qui puissent rendre compte de ces changements systématiques de sens. On peut ainsi citer Pustejovsky (1995) qui propose un mécanisme de *coercition de type*, Nunberg et Zaenen (1997) qui défendent l'idée de *polysémie systématique* (cf. aussi Nunberg, 1995), ou encore la notion, plus complexe, de *facettes*, défendue par Cruse (Cruse, 2000 ; Croft et Cruse, 2004 : chap. 5). Il faut cependant éviter les généralisations trop hâtives et distinguer soigneusement ce qui relève du discursif et qui est effectivement systématique (comme les exemples que nous avons donnés ci-dessus) de ce qui relève de phénomènes lexicaux, qui leur ressemblent à première vue mais qui s'avèrent beaucoup plus rétifs à la généralisation parce que moins systématiques. Il en est ainsi, par exemple, de la *fonction de transfert* postulée par Nunberg et Zaenen (1997) qui régulerait l'emploi d'un nom comptable dans un sens massif. Le passage de comptable à massif ferait appel à un « broyeur universel »[10] : *un lapin*, en devenant *du lapin*, est transformé en « substance lapine » qui peut selon le contexte désigner de la viande de lapin (*J'ai mangé du lapin*), de la fourrure de lapin (*Elle porte du lapin*), ou un mélange indifférencié résultant d'un broyage effectif (*Après que plusieurs camions eurent roulé sur le corps, il y avait du lapin partout sur l'autoroute*). En fait, Kleiber

9. On trouvera chez Victorri (2005) un exemple de méthode de résolution des métonymies intégrées conducteur/véhicule (qui fonctionne d'ailleurs dans les deux sens) dans un système d'extraction d'information.

10. L'expression *universal grinder* a été introduite par Pelletier (1975), selon Kleiber (1999).

(1999 : chap. 4) montre qu'il faut distinguer le dernier exemple (le lapin sur l'autoroute) des deux précédents (le lapin cuisiné et le lapin en manteau). Seul le lapin sur l'autoroute est effectivement le résultat d'un processus systématique (qui porte bien son nom de « broyeur »), applicable à n'importe quelle entité matérielle, mais dans des conditions discursives très contraintes (il faut une situation très particulière, ici la route et les camions, pour que ce sens soit évoqué[11]). En revanche, les acceptations 'viande de lapin' et 'fourrure de lapin' doivent être considérées comme lexicalisées, faisant partie du potentiel sémantique de l'unité lexicale *lapin*, et n'étant pas inférable par une règle générale. En effet, si *de la mirabelle* désigne de l'alcool de mirabelle, *du raisin* ne peut pas dénoter de l'alcool de raisin ou du vin, de même que *de l'orange* n'est pas du jus d'orange, *de l'olive* n'est pas de l'huile d'olive, etc. L'exemple du mot *vison* montre bien d'ailleurs que ces processus sont spécifiques à l'unité considérée : si *du vison* désigne bien de la fourrure de vison, *un vison* dénote plus facilement un manteau qu'un animal, alors que l'hypothèse de fonctions de transfert devrait en faire un sens doublement dérivé (un premier transfert de l'animal « broyé » en fourrure, puis un deuxième transfert, en sens opposé massif → comptable, « découpant » un vêtement dans ladite fourrure…).

D'une manière générale, il faut donc distinguer deux niveaux de modélisation. D'abord, on doit modéliser un processus essentiellement lexical, dans lequel on s'appuie sur les sens conventionnels des unités polysémiques, comme par exemple les sens 'animal' et 'fourrure' pour *vison*, 'fruit' et 'liqueur' pour *mirabelle*, 'animal' et 'viande' pour *poulet*, 'animal', 'viande' et 'fourrure' pour *lapin*, pour sélectionner le sens approprié. Et ce n'est que dans un deuxième temps que les règles discursives doivent être mises en œuvre, si les contraintes imposées par le contexte phrastique l'imposent. Ainsi *du lapin* que l'on mange ou que l'on porte sera analysé comme tel sur une base

11. Le broyage n'est pas le seul mécanisme qui peut opérer la conversion comptable → massif au niveau discursif. Dans *Il y a du sanglier dans cette forêt*, il s'agit, comme le rappelle Kleiber, de sangliers bien entiers ! Citons aussi *Ça, c'est de la belle armoire !* qui n'implique pas, loin de là, que l'armoire en question soit en morceaux.

lexicale, et ce n'est que dans le cas de l'autoroute que le sens 'animal' (dans la mesure où il aura été sélectionné dans le processus précédent), de type comptable, provoquant un conflit avec le partitif *du* qui exige le type massif, conduira à l'utilisation de mécanismes discursifs de changement de type tel que le broyeur universel pour aboutir à l'interprétation pertinente.

Dans cette thèse, nous nous intéresserons essentiellement au premier niveau de modélisation, c'est-à-dire au niveau de l'unité lexicale et de ses sens conventionnels. Il faudra donc que l'on dispose d'un mode de représentation de ces différents sens, or, comme on va maintenant le voir, les linguistes sont très divisés sur ce point.

2.4. Approches de la polysémie lexicale

Deux conceptions radicalement opposées reviennent, d'une certaine manière, à nier la polysémie lexicale, ou, plus précisément, à refuser de lui accorder un statut prééminent en sémantique lexicale. A la suite de Rastier (1994), on peut les appeler la conception nominaliste et la conception essentialiste.

Commençons par la conception nominaliste, dans laquelle se situe la sémantique interprétative de Rastier. Cette conception consiste à dire que le sens lexical « n'est pas doté d'une identité à soi qui définirait un noyau de sens invariant et primordial. Sa définition dépend de conditions objectives telles que le contexte (local puis global) et la situation, mais encore de conditions subjectives qui sont celles de l'interprétation. » (Rastier 1994 : 50). Une autre manière d'exprimer cette idée est de dire que « rien ne peut être représenté en langue qui n'ait auparavant été décrit en contexte » (Rastier 1994 : 62). Dans ce cadre, on conçoit que la polysémie lexicale soit considérée comme un « artefact » dû à des conceptions erronées de la sémantique. Dans un domaine et un contexte donné, les problèmes de polysémie et d'ambiguïtés sont négligeables. Pour Rastier (1994 : 51), « il est bien rare que dans une pratique déterminée on ait à distinguer, parmi les acceptions de *plateau*, un plateau géographique d'un plateau de service, de spectacle, de tourne-disque ou de machine-outil ».

Rastier oppose donc à cette conception la conception essentialiste qui rassemble « diverses théories des stéréotypes, prototypes et archétypes lexicaux qui toutes réifient un noyau de sens infrangible » (Rastier, 1994 : 50). Ce noyau de sens est nommé de différentes manières selon les auteurs : forme schématique (Culioli, 1990 : 115-135), figure morphologique (Pottier, 1987), image schéma (Langacker, 1991), archétype cognitif (Desclés, 1985), motif (Cadiot et Visetti, 2001). La définition que Culioli donne de ses formes schématiques illustre très bien cette notion de noyau de sens : « La forme schématique fournit une configuration abstraite qui, selon les transformations qu'on lui fait subir (translation, décrochage, plongement dans un domaine centré, dans un champ de forces inter-sujets) va modifier la forme (marqueur), sa valeur, sa latitude de co-occurrence » (Culioli, 1990 : 130). On peut rattacher aussi à cette conception les théories qui postulent l'existence pour toute unité d'un « sens de base », appelé aussi « sens premier » ou encore « saisie plénière » (Picoche, 1986), dont tous les autres sens dérivent. C'est différent d'un noyau de sens, dans la mesure où il ne s'agit pas d'une forme abstraite commune à tous les sens mais de l'un des sens de l'unité auquel on fait jouer un rôle privilégié. Néanmoins, il y a la même idée de retrouver par ce biais une univocité de la correspondance entre une forme et un élément de sens. Comme Kleiber le fait remarquer, cette conception essentialiste revient donc, elle aussi, à réfuter l'existence de la polysémie : « Il suffit d'une part, de pouvoir expliquer tous les emplois d'une unité linguistique par un invariant sémantique […] et de constater, de l'autre, que, finalement, l'interprétation de toute unité linguistique se fait par interaction avec les autres éléments contextuels. On est à ce moment-là en droit de conclure au caractère non spécifique du phénomène polysémique, puisque toute unité linguistique se trouve munie d'un seul sens intrinsèque et de sens formés ou déformés en contexte. Autrement dit, toute unité a un sens et toute unité a besoin de contexte pour compléter (ou actualiser) ce sens » (Kleiber, 1999 : 58).

Certes, chacune de ces deux conceptions présente un intérêt théorique indéniable, mais il paraît tout de même très difficile de se passer de la polysémie dès que l'on s'occupe de manière concrète de calcul du sens.

Prenons l'exemple du verbe *monter*. Si l'on se fonde sur une conception nominaliste, ce verbe ne possède pas de noyau de sens. Lorsqu'il est dans l'expression *monter sur un banc*, c'est la présence du nom *banc* qui permet de donner le sens « s'élever pour se trouver plus haut »[12] au verbe *monter*. Il en va de même pour l'expression *monter sur un bureau*. Pourtant, *banc* peut être défini comme « troupe importante d'animaux marins de même espèce et se déplaçant ensemble» dans *banc de poisson*, ou encore « installation servant à la détermination des caractéristiques d'une machine tournante ou d'un appareil fonctionnant à différents régimes» dans *mettre sur un banc d'essai*. Parallèlement, *bureau* peut prendre le sens de « pièce privée ou officielle où l'on effectue un travail de nature plutôt intellectuelle » dans *travailler dans son bureau* et le sens de « organisme ou établissement ouvert au public et dont la vocation est de rendre un service d'intérêt général » dans *aller au bureau de poste* ou *au bureau des objets trouvés*. Comment deux noms, *banc* et *bureau*, pouvant avoir des sens si différents dans d'autres emplois, orientent-ils tous les deux le sens de *monter* vers la même définition « s'élever pour se trouver plus haut » ? Il semble clair que ce qui se joue entre le verbe et les deux noms se situe à un niveau lexical très « basique », indépendamment des différents domaines et contextes, en fait assez variés, dans lesquels ces expressions, *monter sur un banc* et *monter sur un bureau*, peuvent être utilisés. L'hypothèse selon laquelle le sens de *monter* dans les deux énoncés cités proviendrait uniquement du contexte semble donc peu plausible. Et si l'on accepte l'hypothèse selon laquelle *monter* possède ne serait-ce qu'un soupçon de sens lexical, il semble aussi peu plausible qu'il existe autant de *monter* différents qu'il existe de nuances de sens : *monter les marches*, *monter à l'échelle*, *monter sur la montagne*, etc..

[12] Les définitions données dans ce chapitre proviennent du Trésor de la Langue Française informatisé (TLFi).

Nous n'avons volontairement parlé que des sens de *monter* proches d'*escalader*, *gravir*, *grimper*. On retrouvera ce même phénomène dans les autres sens que peut prendre ce verbe tels que *préparer, organiser* (ex. : *monter un projet*) ou encore *augmenter, s'élever* (ex. *la température monte*), etc.[13]… Il y a donc un niveau de description et d'analyse, indépendant du contexte, où l'on peut identifier la polysémie par l'existence de paraphrases différentes. Ainsi, *grand* dans *une grande chambre* et dans *un grand vin* ne se laisse pas paraphraser de la même manière. Dans le premier énoncé, on emploiera plutôt des termes tels que *vaste, spacieux* alors que dans le second énoncé on emploiera *fameux, excellent* (Victorri, 1997 : 41). Et ces opérations de paraphrasage ne sont pas des manipulations réservées aux seuls linguistes (ce qui justifierait la qualification d'« artefact », utilisée par Rastier), mais constituent une opération banale, utilisée par tous les locuteurs de la langue, dans de nombreuses circonstances : apprentissage, explicitation, effets littéraires, et aussi négociation sur le sens des mots (ex : *Tu appelles ça un grand vin, toi ? Je dirais un vin agréable, bon même, mais sans plus...* cf. Victorri, 1997 : 57).

Si l'on adopte la conception essentialiste, on est confronté à un tout autre problème : comment pourrait-on déterminer un noyau de sens de *monter*, qui rende compte à la fois du sens *grimper* et du sens *organiser* ? Ou, si l'on adopte le point de vue d'un sens premier, que l'on associera sans doute au sens *grimper*, comment peut-on dériver le sens *organiser* de ce sens premier ? Dans le cadre de la théorie de Culioli, Paillard considère qu'il faut « déplacer la question de la polysémie vers l'étude de la mise en oeuvre de principes fondamentaux de variation, qui fondent des modes de contribution du lexique à la construction du sens d'un énoncé. Cette variation est indissociable de la mise en place d'un pôle d'invariance définissant l'identité sémantique du mot » (Paillard, 2001 : 101). Il s'agit donc d'un programme de recherche tout à fait cohérent. Mais, face à la « complexité et l'hétérogénéité des facteurs fondant

[13] Nous présenterons une description détaillée des sens de *monter* dans le chapitre 4.

la diversité des valeurs d'un lexème verbal » (Paillard, 2001 : 100), on ne peut remplir cet objectif qu'avec des formes schématiques de verbes qui atteignent un niveau d'abstraction extrêmement élevé afin de convenir à l'ensemble des emplois du verbe. Certes, une forme schématique est par définition abstraite, mais regardons la forme schématique du verbe *jouer* proposée par Romero–Lopez (2002) et reprise par Paillard : « *jouer* signifie qu'une entité **a** est l'actualisateur d'un ensemble **X** de propriétés. Cet ensemble a une cohérence qui lui est propre (indépendamment de toute actualisation). » Le niveau d'abstraction est tel que l'on ne retrouve plus le lien avec le verbe de départ. Cela est sans doute inévitable pour pouvoir rassembler des énoncés tels que *jouer de la guitare, jouer aux courses, jouer les imbéciles, jouer des coudes, la clé joue dans la serrure* sous une seule et même forme schématique, régie par les mêmes principes fondamentaux de variation. Mais du même coup, cela rend illusoire de vouloir utiliser de telles formes schématiques pour calculer effectivement le sens d'un verbe dans un emploi donné.

On doit donc en conclure qu'il ne sert à rien de chercher à tout prix à maintenir l'unicité de représentation des unités lexicales polysémiques. Mais si l'on veut modéliser de façon complète et précise le comportement sémantique de l'unité polysémique considérée, on est rapidement confronté à de nouvelles difficultés : combien doit-on prévoir de sens différents pour un verbe comme *monter* ou *jouer* ? A quel degré de finesse doit-on s'arrêter ? Et comment sélectionner « le » sens pertinent dans un énoncé si l'on multiplie des représentations à la fois concurrentes et très proches ? Comme nous le discuterons en détail plus loin (chap. 4), ce dernier problème se pose de manière d'autant plus aiguë que, suivant les énoncés, une unité peut avoir un sens plus ou moins précis, voire relever de plusieurs sens à la fois. Ces difficultés sont bien sûr le « prix » à payer pour l'abandon d'une représentation sémantique unifiée pour chaque unité polysémique. Nous les aborderons dans le cadre de la présentation du modèle de la polysémie que nous avons choisi, dont la caractéristique essentielle est de proposer une représentation continue des variations de sens d'une unité polysémique.

Comme nous le verrons au chapitre 4, cela permet de rendre compte de l'aspect graduel de ces variations, et donc de pouvoir travailler à différents degrés de finesse en fonction des besoins et des énoncés.

Mais pour le moment, nous allons achever ce premier tour d'horizon en nous concentrant sur les spécificités de la polysémie verbale, qui, comme nous l'avons déjà laissé entendre, semble plus complexe par bien des aspects que la polysémie nominale ou adjectivale.

2.5. Spécificités de la polysémie verbale

Il existe une multitude de modèles proposant une description de la polysémie pour les verbes. Pustejovsky et Busa (1995) parlent de polysémie verbale pour les relations entre les différentes réalisations grammaticales d'un même verbe, par exemple la construction causative (*l'eau a rompu la digue*) et inaccusative (*La digue a rompu*). Copestake et Briscoe (1995) proposent une distinction entre polysémie constructionnelle et extension de sens. La polysémie constructionnelle est une modulation de sens impliquée par le contexte (l'équivalent de ce que Pustejovsky appelle co-composition), alors que l'extension de sens est le processus lexical qui crée des sens dérivés à partir d'un sens premier.

Très tôt, la singularité de la polysémie des verbes a été mise en évidence. Martin (1979) explique que « s'agissant d'un verbe, la complexité du problème s'accroît notablement. Cela tient en grande partie à la double nature de la polysémie verbale : le verbe, en effet, […] peut être touché par la polysémie dans le sémème[14] ; mais il peut l'être aussi dans les actants. L'une est dite interne, l'autre externe. Si la première est proche parente de la polysémie substantive, la seconde apparente le verbe à l'adjectif et présente même des traits qui paraissent n'appartenir qu'au verbe. » (Martin, 1979 : 251)

[14] La notion de sémème recouvre celle de sens et celle d'acception. Dans un dictionnaire, chaque sémème correspond à une définition.

Autrement dit, Martin oppose une polysémie interne, c'est-à-dire qui « naît de la complexité du sémème », à une polysémie externe, c'est-à-dire qui n'est pas propre au verbe mais provient de ses arguments. Martin illustre cette polysémie externe avec le verbe *apprendre*. Dans la construction *apprendre quelque chose à quelqu'un,* ce verbe établit entre un actant, l'agent, et un autre, le destinataire, une relation telle que le destinataire sait ce qu'auparavant il ignorait. Comparons :

(1) apprendre la danse à quelqu'un

(2) apprendre une nouvelle à quelqu'un

Malgré des différences sensibles, comme par exemple le fait que dans (1) le destinataire est plus actif que dans (2), on peut admettre que, pour le moins, l'essentiel de l'opposition est dans la nature de l'objet, et non pas dans la définition de la relation établie entre l'agent et le destinataire. C'est en ce sens que Martin parle de polysémie externe.

Cette distinction entre polysémies interne et externe semble convaincante, mais pour une analyse lexicographique concrète, cela devient très difficile d'établir une distinction claire entre les changement de sens du prédicat et les changement de sens de ses arguments : il est pratiquement impossible de prouver que le sens du verbe ne change pas du tout lorsqu'il apparaît dans des contextes différents. (Stein, 1999 : 113).

Prenons le cas du verbe *changer*. Martin utilise ce verbe pour décrire les variations de sens liées au jeu des prépositions (construction directe, construction indirecte et double construction). Il oppose :

changer quelque chose (*changer les rideaux*)

changer de quelque chose (*changer de rideaux*)

changer quelque chose en *quelque chose* (*changer le plomb en or*)

changer quelque chose contre quelque chose (*changer des florins contre de francs*)

Il distingue alors les sens que prend le verbe dans chacun de ces énoncés. Voici une synthèse des sens de *changer* proposés par Martin, respectivement pour les quatre énoncés précédents (pour plus de détails, voir Martin, 1979 : 254) :

remplacer en gardant des propriétés référentielles communes

remplacer sans garder de propriétés référentielles communes

transformer

échanger

Le problème est que Martin classe ces changements de sens dans la polysémie interne. Pourquoi considérer que les arguments lexicaux sont externes au verbe, et les constructions verbales internes ? De plus, Martin justifie la distinction d'une polysémie des verbes par l'apparition d'une polysémie externe. Or, le fait que l'entourage lexical introduise de la polysémie n'est pas propre au verbe, on retrouve ce phénomène pour les adjectifs et les noms. Selon nous, la singularité de la polysémie des verbes est plus à chercher du côté des constructions verbales[15].

La sémantaxe de J. François propose de décrire cette articulation entre la polysémie verbale et ce qu'il appelle la « polytaxie », c'est-à-dire « l'éventail des co-textes actanciels avec lesquels un verbe (et ses synonymes) se révèle compatible. » (François, 2004). François fait un rapprochement de sa sémantaxe avec d'autres terminologies : les cadres de rôles sémantiques (Chafe 1970, Cook 1989), les structures argumentales (Grishaw 1991), les structures lexico-conceptuelles (Jackendoff 1990), les schèmes (Desclès 1990), les cadres prédicatifs (François 2003a). Dans tous les cas, « la schématisation du co-texte actanciel est conçu comme mettant en jeu des classes syntaxiques, des indications de sous-catégorisation et de restriction de sélection[16] (ou

[15] On retrouve aussi des phénomènes de contraintes de sens par la constructions pour les noms mais de manière beaucoup moins systématique. Véronis (2004) décrit le cas du nom *barrage* : barrage de X par Y ; barrage sur X ; barrage à X.

[16] Nous reviendrons sur la notion de « restriction de sélection » dans le chapitre 3 (cf. § 3.1.3).

« classes d'objets » dans la terminologie développée au LLI) et éventuellement des rôles sémantiques et un classement en termes de caractères aspectuel de la prédication. » (François, 2004). Autrement dit, l'idée est de traiter la polysémie des verbes dans un cadre plus général consistant à traiter en parallèle les phénomènes de changement de construction et les phénomènes de changement de sens.

Pour terminer, nous souhaitons présenter l'approche avec laquelle nous sommes le plus en accord, même si sur certains points nous nous en éloignons puisqu'elle a tendance à minimiser le phénomène de la polysémie. Nous voulons parler des grammaires constructionnelles, introduites par Fillmore *et al.* (1988), dont la pierre d'angle consiste à considérer qu'une construction syntaxique est porteuse d'un sens intrinsèque, indépendamment des unités lexicales qui la composent. Nous développerons plus en détail ce courant dans le chapitre 7 puisqu'il fait partie des piliers théoriques de cette thèse. Contentons-nous uniquement pour l'instant d'en introduire le principe général, et d'indiquer ce qu'il apporte de nouveau à l'analyse de la polysémie verbale.

Goldberg et Jackendoff (2004 : 2) définissent la notion de construction de la manière suivante :

a. Tous les phénomènes linguistiques (associant une forme à un sens) peuvent être placés sur une échelle graduelle allant du plus général (complètement compositionnel) au plus figé (idiosyncratique).

b. Tout sur cette échelle doit être représenté dans un format commun, depuis le plus particulier, comme les unités lexicales prises individuellement, jusqu'au plus général, comme les règles régissant l'ordre des mots. Autrement dit, il ne doit pas y avoir de division de principe entre le lexique et la syntaxe.

c. A différentes positions sur cette échelle, on trouve des éléments de syntaxe associés à un sens conventionnel et en partie idiosyncratique : c'est ce que l'on appelle habituellement des constructions.

Autrement dit, les constructions sont des unités linguistiques à part entière au même titre que les unités lexicales : elles doivent être listées dans le lexique sous la forme d'un couple : une forme et un sens, la forme étant constituée d'un schéma syntaxique et d'éléments de phonologie (quand la construction comportent des mots grammaticaux particuliers). Par exemple, prenons en anglais la *way-construction*, à l'oeuvre dans des énoncés tels que :

(1) *Bill belched his way out of the restaurant* (Bill sortit du restaurant en rotant tout du long)

(2) *Bill slashed his way through the bush* (Bill traversa la brousse à coups de machette)

Elle est définie par le schéma syntaxique :

[SUJ [V *one's way* OBL]]

où SUJ est un syntagme nominal sujet, V un verbe non statique et OBL un syntagme directionnel, et son sens peut être décrit ainsi : l'agent (évoqué par SUJ) se déplace sur une trajectoire (spécifiée par OBL), V évoquant une activité qui accompagne le déplacement (indication de moyen ou de manière).

Selon les auteurs, un des intérêts de cette approche est « une réduction considérable de la polysémie apparente des verbes dans le lexique ». En effet, dans l'énoncé (1), le verbe *belch* (roter) n'est pas converti en verbe de mouvement dans le lexique ou ailleurs. Sa contribution pour le sens de l'énoncé est la même que sa contribution pour le sens de *Bill belched loudly* (Bill rota bruyamment) ; il exprime une fonction physiologique et une émission de bruit. C'est la construction qui est responsable de l'apport au sens global de l'énoncé (1) de la notion de mouvement (ainsi d'ailleurs que celle d'itération pour l'activité qui accompagne le mouvement, quand celle-ci est ponctuelle, comme c'est le cas pour *belch*).

En traitant les constructions comme des unités linguistiques à part entière, les grammaires constructionnelles permettent donc une vision unifiée du calcul du sens :

chaque unité présente dans l'énoncé, qu'il s'agisse du verbe lui-même, de sa construction, ou de ses arguments, apporte sa propre contribution à la construction du sens global de la phrase. En retour, le sens de chacune de ses unités est spécifié, du moins pour celles qui sont polysémiques : notamment pour un verbe polysémique, c'est au cours de ce processus dynamique que l'on doit situer les interactions avec sa construction et ses arguments. Comme on le verra au chapitre 4, c'est sur cette base que nous avons conçu notre modèle de désambiguïsation.

3. Désambiguïsation lexicale automatique

Retrouver le sens d'un mot polysémique dans un énoncé fait partie des principaux verrous du traitement automatique de la langue. En 1949, Weaver est le premier à argumenter de la nécessité de la désambiguïsation lexicale pour la traduction automatique. Aujourd'hui, le nombre d'applications qui ont besoin de désambiguïsation lexicale est très important. Outre la traduction automatique, il y a la recherche d'informations : si l'on recherche dans une base de données conséquente des informations à l'aide d'une requête textuelle, la désambiguïsation permet de donner un sens aux mots de la requête et donc de retourner une information plus ciblée en proposant les informations contenant les mots de la requête mais uniquement dans le sens de la requête, et une information plus complète en proposant des informations ne contenant pas les mots de la requête mais des mots qui ont le même sens que les mots de la requête (De Loupy, 1998). La désambiguïsation est aussi utile pour l'indexation de documents pour les mêmes raisons, c'est-à-dire classer les documents non à l'aide des mots qu'ils contiennent mais à l'aide des sens des mots qu'ils contiennent.

Ces tâches font de plus en plus partie du quotidien notamment avec les bases de données techniques en entreprise, les catalogues de bibliothèques ou tout simplement Internet. Il n'est donc pas surprenant qu'un nombre très important de travaux ait tenté d'automatiser ce processus de désambiguïsation. Si ces travaux sont si nombreux, c'est aussi qu'il n'y a d'accord unanime ni sur la méthode à adopter pour effectuer cette tâche, ni sur les ressources à exploiter.

Plusieurs auteurs ont présenté ces différentes approches (Ide et Véronis 1998, Kilgarriff 1992, Audibert 2003, Mona Talat Diab 2003, Agirre et Martinez, 2001), l'état de l'art le plus complet étant probablement celui de Ide et Véronis, mis à jour en 2003

par Audibert. Il ressort de ces différents états de l'art une distinction implicite ou explicite entre deux grands types d'approches en fonction des ressources utilisées :

Les approches exogènes

Elles regroupent tous les travaux exploitant des ressources extérieures au texte étudié. Ces ressources peuvent être de différentes natures. Certains utilisent des dictionnaires de langue classiques qui ont été mis sous forme électronique, comme par exemple le Robert, ou l'OED pour l'anglais. D'autres utilisent des thésaurus tels que le « Roget's Thesaurus », qui sont sollicités pour leur structure hiérarchisée comme par exemple la description des liens hyperonymiques entre mots. D'autres encore utilisent des dictionnaires particulièrement adaptés au traitement automatique, comme par exemple WordNet qui combine définitions, relations hyperonymiques, relations de synonymie, d'antonymie, etc.[17].

Le principe général de ces approches consiste à utiliser le dictionnaire, ou l'équivalent, comme référence. La désambiguïsation d'un mot polysémique dans un contexte donnée, par exemple une phrase, consiste dans un premier temps à extraire les mots cooccurrents présents dans cette phrase. Parallèlement, pour chaque sens du mot polysémique, on extrait du dictionnaire la liste des mots présents dans chacune des définitions correspondantes. Ensuite, la désambiguïsation consiste à choisir parmi les sens possibles, celui dont la définition possède la liste de mots la plus proche de la liste extraite de la phrase. Lesk (1986) est l'un des premiers à avoir développé cette méthode. Nous allons voir que les modèles ont depuis été largement optimisés, mais le principe reste le même.

[17] Le dictionnaire électronique WordNet sera cité plusieurs fois dans cette partie. Nous en proposons une présentation ainsi qu'un extrait en annexe F.

Les approches endogènes

Par opposition, nous appelons approches endogènes tous les travaux dont le processus de désambiguïsation ne fait appel à aucune source extérieure. Le principe général consiste à désambiguïser un mot polysémique à l'aide de calculs statistiques sur les instances de ce mot dans le corpus et les différents contextes dans lequel il est employé. L'objectif est ensuite d'établir des corrélations entre les différents emplois du mot polysémique et les sens de ce mot. L'une des principales tâches de ces approches consiste alors à rechercher, parmi les relations de cooccurrences, les relations syntaxiques, les relations thématiques, quel est le critère pertinent, ou plus raisonnablement, quels sont les critères pertinents pour la désambiguïsation. L'autre difficulté étant de déterminer comment combiner ces différents critères.

Nous appellerons mixtes, les approches combinant ressources endogènes et exogènes.

Plan de présentation

Dans cette partie, nous allons essayer à notre tour de dresser un panorama des travaux de désambiguïsation automatique. Comme nous venons de le dire, il y a une grande hétérogénéité dans ces travaux, tant du point de vue des a priori théoriques, que des méthodes, des outils ou encore des objectifs visés. En conséquence, nous n'allons pas essayer de faire un état de l'art exhaustif, nous allons plutôt nous focaliser sur des questions et des approches particulières ayant un lien avec notre propre travail. Pour présenter ces approches, nous avons choisi six travaux, chacun ayant en toile de fond une théorie ou un modèle ayant fait l'objet de plusieurs travaux.

Nous commencerons par deux approches assez éloignées de notre étude de par les résultats attendus puisqu'elles présentent un modèle de désambiguïsation pour un nombre limité de mots : Yarowsky (2000) permettra de présenter un système d'apprentissage supervisé et Peng, Ito et Furugori (2001) illustrera l'utilisation des réseaux sémantiques. Les quatre travaux suivant cherchent à désambiguïser n'importe

quel mot plein (*content word*). Wilks et Stevenson (1998) est un exemple remarquable des approches mixtes. Ils présentent aussi une méthode élégante de combinaison de différentes sources de connaissances. Les trois dernières études présentent des points de similitude importants avec notre approche : Inkpen et Hirst (2003) utilisent un dictionnaire de synonymes, Resnik (1995) propose de calculer des classes de sens, et Schütze (1998) propose de représenter les mots dans un espace vectoriel muni de la distance utilisée par l'analyse sémantique latente (en anglais, LSA : *Latent Semantic Analysis*).

Une fois ces six travaux détaillés, nous présenterons certains problèmes récurrents dans les modèles de désambiguïsation, notamment concernant l'évaluation ou encore la qualité des ressources utilisées. Enfin, nous terminerons en justifiant notre intérêt particulier pour deux sources de connaissance : les relations syntaxiques et les classes sémantiques.

3.1. Six travaux représentatifs

3.1.1 Yarowsky (2000)

Objectif et méthode

L'objectif est de désambiguïser le sens d'un nombre limité de mots dans un corpus donné. Ce travail est un exemple type de l'utilisation de l'apprentissage supervisé sur corpus. Il illustre aussi de ce fait un cas d'approche endogène. Le principe de l'apprentissage supervisé consiste en deux étapes : une étape d'apprentissage sur une partie du corpus, et une étape de désambiguïsation sur le reste du corpus. C'est-à-dire qu'une fois l'étape d'apprentissage effectuée, on doit être capable de déterminer automatiquement le bon sens de chacune des occurrences des mots polysémiques étudiés sur le reste du corpus.

Ici, l'apprentissage se fait sur une liste de décision hiérarchique. Une liste de décision (Rivest, 1987) est une séquence ordonnée de paires "attribut-valeur" décrivant les objets que l'on désire classifier. On peut interpréter les listes de décisions comme

une séquence ordonnée de règles du type “si ... alors”. La classification de cette liste est effectuée en testant chaque règle sur le corpus d’apprentissage. Lors de la désambiguïsation, le sens donné à l’occurrence du mot polysémique sera celui associé à la première règle de la liste s’appliquant à cet exemple. Lorsque aucune règle ne s’applique, une valeur par défaut est retournée. Yarowsky insiste sur le fait que ses listes de décision sont hiérarchiques, c'est-à-dire qu’une règle appartenant à une première liste de décision peut pointer sur un sens comme dans toute liste de décision classique, mais elle peut aussi pointer sur une autre liste de décision comme on peut le voir dans le figure 3.1 où la liste nommée *promise* peut pointer sur la liste *promise.LN* qui elle-même peut pointer sur la liste *promise.L4*. L’intérêt d’une telle structure est de pouvoir notamment construire des règles conditionnelles de type « si x et y alors z ».

Technique

Les décisions sont prises en fonction de la présence ou non de certains patterns. Chaque pattern étant composé de trois attributs. Le premier, nommé *Loc* dans la figure 3.1, concerne la localisation du pattern. *Loc* peut être égal à 0 si le pattern concerne les caractéristiques du mot vedette lui-même, égal à -2,-1,+1,+2 s’il concerne les caractéristiques d’un pattern voisin du mot vedette, ou plus généralement $\pm k$ pour des patterns se trouvant dans une fenêtre de mots délimitée. Lorsque le mot étudié est un nom, *Loc* peut aussi correspondre à une relation syntaxique : V/OBJ (le verbe dont le mot vedette est objet), SUBJ/V et MODNOUN (tête nominale modifiée par le mot vedette)[18]. Le deuxième attribut d’un pattern, nommé *type* dans la figure 3.1, correspond au type de pattern. Il y a cinq types possibles : W = mot littéral, L = lemme, P = catégorie grammaticale, C = classe de mots (par exemple « noms de pays ») et Q = questions (par exemple : est-ce que le mot est écrit en capitales ?). Le troisième attribut correspond à la valeur du type, nommé *token* dans la figure 3.1.

[18] Voir Dini, Tomaso et Segond (1998) pour une utilisation similaire des relations syntaxiques dans une approche non-supervisée.

Table I. Partial decision list hierarchy for the SENSEVAL word **promise**

Top-level Decision List for **promise**

Pattern Loc	Typ	Token		Next List	Empirical Sense Distribution **1**	**3**	**4**	**4.1**	**4.2**	**4.3**	**4.4**	**5**	**6**
+0	P	NOUN	→	$LN_{(\Downarrow)}$	0	0	297	53	5	37	11	22	93
+0	P	VERB	→	LV	440	115	0	0	0	0	0	0	0

⇓

Mid-level Decision List for **promise.LN** (noun)

Pattern Loc	Typ	Token		Next List	Empirical Sense Distribution **4**	**4.1**	**4.2**	**4.3**	**4.4**	**5**	**6**
V/obj	L	keep/V	→	**4.3**	0	0	0	31	0	0	0
V/obj	L	break/V	→	**4.4**	0	0	0	0	11	0	0
V/obj	L	make/V	→	L1	2	44	0	0	0	0	2
V/obj	L	give/V	→	L2	0	0	5	1	0	1	2
+0	W	promises	→	L3	115	5	0	0	0	0	1
+0	W	promise	→	$L4_{(\Downarrow)}$	180	3	0	1	0	21	88

⇓

(Abbreviated) Terminal Decision List for **promise.L4** (promise-noun-singular)

Pattern Loc	Typ	Token		Output Sense	LogL	Empirical Sense Distribution **4**	**4.1**	**4.2**	**4.3**	**4.4**	**5**	**6**
+1	W	to	→	**4**	9.51	41	0	0	0	0	0	0
-1	W	of	→	**6**	8.16	0	0	0	0	0	0	12
-1	L	early/J	→	**6**	7.38	0	0	0	0	0	0	7
V/obj	L	show/V	→	**6**	7.27	0	0	0	0	0	0	13
+1	W	at	→	**6**	6.16	0	0	0	0	0	0	3
-1	L	firm/J	→	**4**	5.74	6	0	0	0	0	0	0
+1	L	do/V	→	**4**	5.70	3	0	0	0	0	0	0
-1	W	such	→	**6**	5.57	0	0	0	0	0	0	2
-1	W	much	→	**6**	5.57	0	0	0	0	0	0	2
+1	W	when	→	**6**	5.57	0	0	0	0	0	0	2
+1	W	on	→	**6**	5.57	0	0	0	0	0	0	2
+1	W	as	→	**6**	5.57	0	0	0	0	0	0	2
-1	W	your	→	**4**	5.16	2	0	0	0	0	0	0
+1	W	during	→	**4**	5.16	2	0	0	0	0	0	0
$\pm k$	L	free/J	→	**4**	4.74	15	0	0	0	0	0	0
V/obj	L	trust/V	→	**4**	4.74	3	0	0	0	0	0	0
$\pm k$	L	support/N	→	**4**	4.64	14	0	0	0	0	0	0
$\pm k$	L	election/N	→	**4**	4.29	11	0	0	0	0	0	0
subj/V	L	contain/V	→	**4**	4.18	2	0	0	0	0	0	0
V/obj	L	win/V	→	**4**	4.16	2	0	0	0	0	0	0
V/obj	L	repeat/V	→	**4**	4.16	2	0	0	0	0	0	0
V/obj	L	honour/V	→	**4**	4.16	2	0	0	0	0	0	0
-1	L	rhetorical/J	→	**5**	4.09	0	0	0	0	0	1	0
-1	L	increase/V	→	**5**	4.09	0	0	0	0	0	1	0
-1	L	future/J	→	**5**	4.09	0	0	0	0	0	1	0

Figure 3.1 : Arbre hiérarchique de décision (extrait de Yarowsky 2000)

Algorithme

L'apprentissage supervisé va consister à pondérer chacune des propriétés décrites afin de générer une liste de décision hiérarchisée. Pour cela, les règles sont

ordonnées par valeur de log-vraisemblance. C'est-à-dire que pour chaque sens s_j du mot ambigu et chaque pattern c_i tiré de l'ensemble d'apprentissage, on calcule :

$$\left| \log\left(\frac{P(s_j|c_i)}{P(\neg s_j|c_i)} \right) \right|$$

Les probabilités étant dérivées des fréquences relatives empiriques des événements dans le corpus d'apprentissage. Les patterns les plus fortement reliés à un sens obtiennent les valeurs les plus élevées. La liste de décision est alors créée en ordonnant les paires sens/pattern sur la base de la log-vraisemblance. Si aucune règle ne s'applique, le sens le plus fréquent est assigné. La figure 3.1 décrit une partie de la liste de décision obtenue par cet algorithme pour le mot *promise*.

Résultats

Les résultats présentés sont ceux obtenus lors de la campagne d'évaluation SENSEVAL de 1998. Cette étude a été classée première de tous les systèmes supervisés avec une précision moyenne de 78,9%. Regardons plus en détail pour chaque type de catégorie grammaticale. Les taux de précision pour les adjectifs, les noms, les adverbes et les verbes sont respectivement de 77,3%, 87,0%, 78,1% et 74,3%. Ces chiffres peuvent être comparés aux précisions moyennes pour les mêmes catégories grammaticales, tous systèmes confondus : 72,7%, 81,7%, 73,7% et 66,4%. On peut noter que pour tous les systèmes, les verbes sont les plus difficiles à désambiguïser. C'est aussi sur cette catégorie que le modèle de Yarowsky fait la différence. Cela peut s'expliquer par l'introduction de relations syntaxiques dans son système de règles.

Yarowsky montre aussi que la structure hiérarchique de la liste de décision est déterminante pour l'efficacité de son modèle puisque si les listes proposées sont mises « à plat », la moyenne de précision chute de 7,3%.

Discussion

Les résultats obtenus dans cette étude sont impressionnants. Nous émettons toutefois une réserve concernant l'approche supervisée. Elle soulève plusieurs points à discuter : le premier concerne le faible nombre de corpus étiquetés sémantiquement. Ce manque s'explique par le coût considérable de construction d'un tel corpus. Ng (1997) a estimé à 16 années-humains (c'est-à-dire l'équivalent du travail de 16 humains pendant un an) l'effort à fournir pour obtenir des données d'apprentissage significatives pour 3 200 mots anglais. Mihalcea et Chklovski (2003) proposent une estimation similaire avec 80 années-humains pour 20 000 mots. Or, ces corpus sont indispensables pour la phase d'apprentissage et ils impliquent un étiquetage à la main qui soulève différentes questions[19] : quel dictionnaire de référence choisir et pourquoi ? Comment prendre en compte les « désaccords » inter juges s'il y a plusieurs annotateurs ? Enfin, peut-on raisonnablement désambiguïser n'importe quel mot avec cette méthode ? Nous reviendrons sur le jugement des annotateurs dans le paragraphe 5.1[20], mais nous pouvons déjà répondre par la négative à la dernière question puisque l'extrait de la liste de décision de *promise* présentée par Yarowsky nous montre le travail préalable important qui doit être effectué manuellement sur chaque mot pour obtenir une désambiguïsation correcte. Par exemple les collocations idiomatiques telles que *keep a promise* (respecter une promesse) doivent être identifiées pour chaque nouveau mot à désambiguïser.

3.1.2 Peng Ito et Furugori (2001)

Objectif et méthode

Ici, l'objectif est aussi de désambiguïser un nombre de mots restreints dans un texte tout venant mais cette fois l'étude est basée sur une approche exogène. Le principe

[19] De nombreux travaux tentent de réduire la taille du corpus d'apprentissage à l'aide de différentes méthodes comme l'algorithme de *boosting* (Escudero, Martinez et Rigau, 2000).

[20] Voir Véronis (1998, 2004) et Habert *et al.* (1997) pour une discussion détaillée sur ces questions.

général de l'approche exogène consiste à confronter le mot à désambiguïser et son contexte d'un côté et la définition de ce même mot dans un dictionnaire de l'autre. Avec toujours cette idée qu'un mot polysémique dans un contexte donné peut être désambiguïsé en sélectionnant parmi ses sens celui dont la définition du dictionnaire possède le plus de mots en commun avec les mots présents dans le contexte (Lesk, 1986). Cette méthode est très sensible aux mots qui se trouvent dans chaque définition : la présence ou l'absence d'un mot donné peut radicalement changer le résultat. Dans cette étude, Peng *et al.* proposent de réduire ce problème en utilisant un réseau sémantique. Ainsi un même nœud du réseau peut être activé par différentes sources. On ne va plus confronter deux listes de mots mais deux zones dans le réseau sémantique, ces deux zones ayant été activées par les deux listes de mots initiales.

Les réseaux apparaissent en sémantique avec la description de la notion « d'amorçage sémantique » (*semantic priming*). Cette notion est basée sur l'hypothèse selon laquelle l'introduction d'un concept dans un énoncé va influencer et faciliter la compréhension de concepts ultérieurs sémantiquement reliés (Meyer et Schvaneveldt, 1971). Les réseaux sémantiques sont exploités selon des modèles de « propagation d'activation » (*spreading activation*), c'est-à-dire que dans un réseau sémantique, les concepts sont activés lorsque les mots correspondants sont mentionnés dans le document, et cette activation est transmise aux noeuds qui sont connectés à ces concepts. Il est donc possible qu'un même noeud soit activé par différentes sources. Bookman (1987) a été l'un des premiers à utiliser ce type de réseaux pour la désambiguïsation.

Dans de nombreux modèles utilisant des réseaux sémantiques, notamment Hiro *et al.* (1996), le sens d'un mot *w* est représenté dans le réseau par les nœuds activés en fonction des mots présents dans la définition de *w*. Ici, les auteurs proposent une autre solution avec pour objectif de pouvoir comparer du contexte à du contexte et non à des définitions. Ainsi, ils extraient du Web les phrases qui contiennent le mot polysémique à désambiguïser. Ils classent manuellement les phrases en fonction du sens du mot

polysémique et sélectionnent six instances pour chaque sens. Ce sont ces paquets de six phrases qui remplaceront les définitions. Nous supposons que ce sont des sens décrits dans un dictionnaire mais cela n'est pas précisé explicitement.

Construction du réseau et algorithme de désambiguïsation

Commençons par la construction du réseau. Ce réseau est construit à partir du corpus EDR[21] (corpus Electronic Dictionary Research contenant 160 000 phrases). Il contient 1 845 nœuds soit tous les mots lexicaux (*content words*) dont le nombre d'occurrences dans le corpus est supérieur à 60. Chaque nœud possède 100 liens, correspondant aux 100 mots les plus cooccurrents. Le degré de cooccurrence entre deux mots w_1 et w_2 se calcule à l'aide de l'information mutuelle :

$$I(w_1, w_2) = \log_2\left(\frac{N_1 * f(w_1, w_2)}{f(w_1) f(w_2)}\right)$$

N_1 étant le nombre total de phrases et *f(x)* la fréquence relative de *x*.

L'information mutuelle entre deux mots w_1 et w_2 consiste à comparer le nombre de fois que ces deux mots cooccurrent, ici cela correspond à la présence de w_1 et w_2 dans une même phrase, avec le nombre de fois qu'ils devraient cooccurrer théoriquement en fonction de la fréquence respective de w_1 et w_2. Plus la fréquence réelle est supérieure à la fréquence théorique, plus les deux mots sont associés.

Le principe consiste ensuite à utiliser le réseau construit comme base de référence pour comparer d'un côté le contexte du mot *w* à désambiguïser et de l'autre chacun des sens du mot *w*. Voici les quatre étapes de l'algorithme :

(a) activer le réseau sémantique (CSN) en utilisant les six phrases extraites manuellement du Web et fournir le vecteur des nœuds, pour chaque sens du mot polysémique *w*.

[21] Site de l'institut Japan Electronic Dictionary Research Institute, qui a développé le corpus EDR : http://www.iijnet.or.jp/edr/

(b) activer le réseau sémantique (CSN) en utilisant une portion du texte dans lequel *w* apparaît (50 mots avant, 50 mots après) et fournir le vecteur des nœuds V.

(c) Calculer la similarité entre V et chaque vecteur de nœuds calculés en (a)

(d) Sélectionner le sens qui maximise la similarité comme le sens de w dans le texte.

Décrivons le processus d'activation : l'activation est diffusée lorsqu'un mot w_k dans une phrase ou dans le contexte du mot polysémique *w* correspond au mot du nœud n_k du réseau. On utilise la formule suivante pour calculer le potentiel a_i, du nœud n_i (i=1,…,1 845) au moment t+1 :

$$a_i(t+1) = \begin{cases} a_i(t) + I(n_i, n_k) & \textit{si } n_i \textit{ et } n_k \textit{ ont un lien dans le CSN} \\ a_i(t) & \textit{sinon} \end{cases}$$

où $I(n_i,n_k)$ est l'information mutuelle entre w_i et w_k.

Résultats

La figure 3.2 décrit les résultats obtenus pour les dix mots polysémiques étudiés, avec leurs sens, le nombre d'instances et les instances identifiées correctement par l'algorithme.

Table 1: Polysemous Words Tested

Words	Senses	Instances	Resolved(%)
band	group of musicians	19	18 (94.7)
	strip or stripe	12	11 (91.7)
cabinet	administrative organ	24	23 (95.8)
	shelf	17	16 (94.1)
court	judicial	163	153(93.9)
	area for ball game	19	18 (94.7)
crane	machine	16	14 (87.5)
	bird	21	20 (95.2)
palm	tree	20	19 (95.0)
	hand	52	49 (94.2)
plant	living thing	86	79 (91.9)
	factory	25	19 (76.0)
sentence	group of words	41	36 (87.8)
	punishment	67	61 (91.0)
slug	bullet	26	24 (92.3)
	animal	16	14 (87.5)
tank	combat vehicle	13	11 (84.6)
	water-filled place	13	13 (100)
trial	action of judging	89	82 (92.1)
	test	17	16 (94.1)

Figure 3.2 : Résultats obtenus sur les dix mots testés

On peut noter tout de même que tous les mots testés relèvent plus de l'homonymie que de la polysémie. Par exemple, le mot *palm,* qui est détaillé dans leur étude ne possède selon eux que deux sens nominaux : *arbre* et *main*. Or, WordNet énumère déjà quatre sens : *arbre*, *main* mais aussi *médaille* et *unité de mesure*. A partir de là, le taux de désambiguïsation annoncé, 92,1%, devient beaucoup moins spectaculaire et de nombreuses méthodes peuvent arriver à ce même succès.

Discussion

Ce travail propose une description claire d'un moyen d'exploiter un réseau sémantique pour la désambiguïsation. Le fait que le réseau soit construit à l'aide d'un

corpus semble mieux adapté pour représenter le vecteur d'un contexte extrait d'un texte qu'une structure artificielle provenant d'un dictionnaire particulier.

Un autre point intéressant pour notre approche est la confrontation du contexte du mot à désambiguïser avec d'autres phrases. Néanmoins, la méthode utilisée pour extraire manuellement les phrases de référence ainsi que le choix du nombre de sens limité à deux reste très discutable.

3.1.3 Wilks et Stevenson (1998)

Objectif et méthode

Nous passons maintenant à des études plus proches de nos travaux puisque l'objectif est de désambiguïser n'importe quel mot plein (*content word*). L'intérêt de l'étude de Wilks et Stevenson est aussi de présenter un cas d'approche mixte en combinant dictionnaire et apprentissage supervisé sur corpus. Les sens recherchés sont ceux donnés par le *Longman Dictionary of Contemporary English* (LDOCE).

Plusieurs sources de connaissances sont exploitées dans cette étude. D'abord un étiqueteur morphosyntaxique (*tagger*) est utilisé pour éliminer les sens qui correspondent à d'autres catégories syntaxiques que celle donnée par l'étiqueteur. Ensuite, trois sources de connaissances de nature différente, toutes les trois issues du LDOCE, sont utilisées. La première utilise de manière classique les définitions du dictionnaire avec les mots qu'elles contiennent. La deuxième utilise les indications de domaine et de registre de langue (appelées *pragmatic codes*) présentes dans la plupart des définitions. Enfin la troisième utilise les restrictions de sélection qui limitent à certaines classes sémantiques (humain, animal, plante, objet solide, etc.) les arguments d'un verbe ou les noms qu'un adjectif modifie (nous reviendrons à plusieurs reprise sur la description des restrictions de sélection dans ce chapitre). Ce qui fait aussi l'originalité de leur méthode est d'appliquer un algorithme d'apprentissage pour déterminer la combinaison optimale de ces trois sources de connaissances pour un mot donné.

Utilisation des différentes sources de connaissance

A partir de chacune des trois sources de connaissance est construit un module qui propose un sens ou une liste ordonnée de sens pour chaque mot analysé.

En ce qui concerne les définitions de dictionnaire, le principe reste classique, c'est-à-dire que le sens d'un mot polysémique w qui est privilégié est celui dont la définition possède le plus de mots en commun avec le contexte d'emploi (la notion de contexte n'est pas plus précisée). La différence ici est que l'on ne choisit pas un seul sens, mais on suggère un ensemble de sens. De plus ce module est optimisé en pondérant la contribution de chaque mot d'une définition par le nombre de mots utilisés dans cette définition.

En ce qui concerne les indications de domaine (les « codes pragmatiques »), l'idée est d'utiliser la cohérence thématique des textes. Pour cela les auteurs utilisent la hiérarchie des codes pragmatiques du dictionnaire LDOCE. L'objectif est d'optimiser le nombre de codes pragmatiques du même type dans la partie du texte contenant le mot analysé. Pour cela ils confrontent uniquement les codes pragmatiques des noms, mais sur l'ensemble du paragraphe contenant la phrase.

Enfin pour les restrictions de sélection, elles sont utilisées suivant les principes de la sémantique préférentielle de Wilks (1975). Les noms sont marqués directement par leurs traits sémantiques (humain, animé, abstrait, etc.). Les verbes sont marqués par les propriétés sémantiques de leurs arguments (sujet, objet, etc.), et les adjectifs, par les propriétés du nom qu'ils modifient. C'est ce qui est appelé, pour les verbes et adjectifs des restrictions de sélection. Lorsqu'un mot possède plusieurs sens, chacun de ces sens possède ses propres traits sémantiques (pour les noms) ou ses propres restrictions de sélections (pour les verbes et adjectifs). Ainsi, l'ensemble de sens de mots possibles pour qu'une phrase respecte les restrictions de sélection est limité. C'est ce principe qui est utilisé ici.

Algorithme d'apprentissage

Le calcul de désambiguïsation va consister à combiner les trois modules décrits ci-dessus. L'algorithme est basé sur un apprentissage supervisé qui conduit à produire une « liste de décision », c'est-à-dire un ensemble ordonné de règles de décision qui opèrent sur les sorties des trois modules (comme nous l'avons déjà vu pour Yarowsky 2000). Voici un exemple d'une telle règle de décision : « Si l'unité est un nom, et que le module utilisant les codes pragmatiques a retourné une valeur unique avec un degré de fiabilité élevé, alors choisir le sens correspondant ».

Résultats

Les auteurs ont testé ce système sur une portion de texte extraite du corpus SEMCOR, qui était au départ étiqueté sémantiquement par les synsets de WordNet, que les auteurs ont pu « traduire », en sens du LDOCE en utilisant une ressource appropriée. Le test a porté sur 2 021 mots étiquetés avec les sens de LDOCE (la portion de texte contenant en tout 12 208 mots). Les 2 021 occurrences renvoient à 1 068 mots différents avec une polysémie moyenne de 7,65. Pour évaluer leurs résultats, il proposent comme borne inférieure (*baseline*) le pourcentage de mots correctement étiquetés en choisissant le premier sens de chacun, ce qui donne un taux de 49,8% de désambiguïsation correcte. L'apprentissage de la liste de décision se fait sur 1 861 occurrences et les 200 restants sont réservés pour l'évaluation. Leur modèle a un taux de réussite de 70% pour retrouver le premier sens étiqueté, et un taux de réussite de 83,4% pour retrouver le sens étiqueté parmi la liste de sens proposés par leur modèle. Un simple système de « vote majoritaire » entre les trois modules ne donne que 59% de taux de réussite, ce qui met en valeur tout l'intérêt de la liste de décision.

Discussion

Le fait que cette étude soit basée sur un apprentissage supervisé l'éloigne de notre approche. Cependant il est intéressant de noter d'une part que les auteurs cherchent à désambiguïser n'importe quel mot, et surtout qu'ils utilisent de nombreuses

sources de connaissance, provenant de ressources différentes. Autrement dit, l'apprentissage ne se fait pas seulement sur corpus mais aussi à partir de critères extraits d'un dictionnaire. Leur taux de réussite (83,4%) pour une désambiguïsation « grossière » est très bon. Ils améliorent encore leur système (Stevenson et Wilks, 2001) en combinant les sources de connaissances suivantes (taux de réussite de 90%) :

– filtrage basé sur l'étiquetage morphosyntaxique

– collocations générées à partir du corpus

– chevauchement avec les définitions du dictionnaire LDOCE

– catégories de sujets

– restrictions de sélection

3.1.4 Inkpen et Hirst (2003)

Les trois études qui suivent correspondent au troisième type que nous avions évoqué en introduction. Ce sont des études qui traitent de désambiguïsation mais avec des objectifs qui diffèrent légèrement. Par exemple pour Inkpen et Hirst (2003) l'objectif est de désambiguïser un mot en fonction du groupe de synonymes dans lequel il se trouve. Nous avons tenu à présenter ces trois études parce qu'elles utilisent des ressources ou des méthodes proches des nôtres.

Inkpen et Hirst travaillent sur les dictionnaires *Webster's dictionnary of synonyms* (Gove 1984) et *Choose the right word* (CTRW, Hayakawa 1994). Dans ces dictionnaires, une entrée contient un groupe (*cluster*) de synonymes, une définition décrivant le cœur de sens qui réunit ces synonymes et une description des différences entre ces synonymes. Ces différences incluent des nuances de sens, de style ou d'emploi. Un exemple d'une partie d'une entrée du CTRW est présenté en figure 3.3.

Cluster: **acumen, acuity, insight, perception**
These nouns all refer to a highly developed mental ability to see or understand what is not obvious. **Acumen** has to do with keenness of intellect and implies an uncommon quickness and discrimination of mind. It requires acumen to solve an intricate problem in human relationships, or to emerge unscathed from a venture into penny stocks. . . . . .
Acuity means sharpness or keenness, and is applied exclusively to perception: visual acuity; The intelligence test was used as a basis for judging the applicant's mental acuity. *See* KEEN, SENSATION, VISION, WISDOM. *Antonyms*: bluntness, dullness, obtuseness, stupidity.

Figure 3.3 : Exemple d'entrée du dictionnaire CTRW

L'objectif est alors de désambiguïser le sens des synonymes pour chaque entrée. Cette étude s'inscrit dans le cadre d'un grand projet dont l'objectif est d'obtenir automatiquement des connaissances sur les synonymes et sur leurs différences à partir d'un dictionnaire tel que le CTRW. Les auteurs insistent sur l'intérêt d'un tel système pour la traduction automatique, c'est-à-dire pour traduire un mot par le bon synonyme, en gardant le sens de la phrase et surtout en gardant les nuances de sens impliquées.

Méthode

Ils présentent un algorithme de désambiguïsation qui utilise le fait que chaque synonyme d'un cluster peut aider à désambiguïser les autres, et le fait que le texte de description du « cœur » de sens d'un cluster est un contexte riche pour la désambiguïsation.

La désambiguïsation d'un mot *w* appartenant à une entrée du dictionnaire CTRW va consister à confronter le contenu de cette entrée avec le contenu de chaque définition de sens dans WordNet. Par exemple *acumen* a deux sens dans WordNet acumen#n#1 décrit par « *a tapering* » (aiguisement) et acumen#n#2 décrit par « *shrewdness shown by keen insight* » (perspicacité). L'idée est d'être capable de dire que dans l'entrée présentée figure 3.3, seul le second est pertinent. Les auteurs précisent que plusieurs sens peuvent être pertinent pour un même mot, c'est-à-dire qu'on peut rapporter le problème à un ensemble de décisions binaires : pour chaque sens du mot, on décide s'il est pertinent ou non.

Technique

Commençons par décrire les différents paramètres pris en compte pour la désambiguïsation.

- ***Intersection entre le texte et les définitions***

Le principal indicateur pour les sens pertinents est la taille de l'intersection entre le texte de l'entrée et les définitions de sens de WordNet. En cela, c'est une approche du type de Lesk (1986). Lors du calcul, le mot à désambiguïser et les mots outils (déterminants, conjonctions, etc.) sont exclus. Avec ce procédé, il peut arriver qu'un texte et une définition soient rapprochés uniquement parce qu'ils ont des mots courants en commun. Pour pallier ce problème, Inkpen et Hirst proposent d'attribuer un « score d'importance » à chaque mot de l'intersection : le score pour le mot *i* dans l'entrée *j* est $tf \cdot idf_{i,j} = n_{i,j} \log \frac{n_i}{N}$ (Salton, 1989), où $n_{i,,j}$ est le nombre d'occurrences du mot *i* dans l'entrée *j*, n_i est le nombre d'entrées qui contiennent le mot *i*, et *N* est le nombre total d'entrées. Le seuil choisi pour déterminer si le sens est pertinent ou non est calculé par apprentissage d'un arbre de décision.

Le même procédé est utilisé pour calculer l'intersection du texte de l'entrée avec la définition des mots « reliés » au mot à désambiguïser, c'est-à-dire les hyperonymes, hyponymes, méronymes, etc.

- ***Les autres mots des synsets faisant partie des synonymes du cluster.***

Par exemple si le cluster est *afraid, aghast, alarmed, anxious, apprehensive, fearful, frightened, scared*, lorsque l'on examine le sens de *anxious*, le sens correspondant au synset *anxious#n#1, apprenhensive#a#2* est pertinent parce que l'autre mot du synset est *apprehensive* et qu'il fait partie des mots du cluster.

- ***Les antonymes***

Comparaison des antonymes du CTRW avec ceux de WordNet, avec l'idée que si deux mots véhiculent les mêmes antonymes, ils auront tendance à être synonymes.

- ***Vecteurs de contexte***

Il arrive que l'intersection entre texte et définition soit vide alors qu'ils sont sémantiquement proches. Par exemple, le sens *reserved*#a#2 avec la définition « *marked by self-restraint and reticence* » de WordNet et le texte décrivant le cluster *aloof, detached, reserved* ont une intersection vide. Inkpen et Hirst proposent d'utiliser des cooccurrences de second ordre à la manière de Schütze (1998).

Algorithme : utilisation d'un arbre de décision

L'arbre de décision est utilisé pour déterminer la meilleure combinaison des indicateurs. Ils utilisent l'algorithme C4.5 (algorithme de Quinlan[22]) sur 904 clusters. Les attributs pour chaque cluster sont les valeurs des indicateurs décrits précédemment :

- intersection entre texte du cluster et définition (valeur numérique)
- intersection entre texte du cluster et définition pour les mots liés (valeur numérique)
- mots appartenant au synset (0 ou 1)
- mots appartenant au synset pour les mots liés (0 ou 1)
- antonymes (0 ou 1)
- cosinus entre les vecteurs de contexte (valeur numérique)

La sortie de l'algorithme est une classification binaire : le sens s_k est pertinent ou non.

[22] cf l'adresse suivante pour une description de l'algorithme
http://www.grappa.univ-lille3.fr/~gilleron/PolyApp/node13.html

Résultats

Pour l'évaluation, les résultats obtenus par l'algorithme sont confrontés aux choix faits par six évaluateurs. Pour cela, 50 des 914 clusters sont extrait aléatoirement du dictionnaire. Ces 50 clusters contiennent 282 synonymes avec 904 sens au total. A partir de l'entrée du CTRW et son contenu, l'évaluateur doit décider pour chaque sens, s'il est pertinent ou non. Sur les 904 décisions, 584 ont été votées à l'unanimité par les 6 évaluateurs. 156 ont été votées à 5 contre 1, 108 ont été votées à 4 contre 2 et 56 à 3 contre 3. Cela donne un accord inter évaluateurs de 85 %.

La solution de référence pour l'évaluation est le choix majoritaire des juges. Pour les 56 sens indéterminés, les auteurs on tranché au cas par cas. Ainsi, le taux de précision obtenu est de 82,5%, c'est à dire une précision très proche de l'accord interjuge. Un autre résultat intéressant est que la précision obtenue pour les verbes est meilleure que pour les noms et les adjectifs, respectivement 83,2%, 81,8% et 78,3%.

Discussion

La précision obtenue dans cette étude est assez convaincante. Néanmoins la singularité de l'étude rend difficile la comparaison des résultats obtenus. Les auteurs admettent d'ailleurs que la tache de désambiguïsation des synonymes d'un cluster est nettement plus facile qu'une désambiguïsation classique étant donné que le contexte exploité est une définition et non un contexte classique. Notons aussi que le taux de polysémie du CTRW est assez faible (3,18) ce qui facilite encore la tâche. Cette étude est néanmoins intéressante par son utilisation d'un dictionnaire de synonymes. Indirectement, elle montre que l'utilisation des synonymes est un bon moyen de désambiguïser un mot.

3.1.5 Resnik (1995)

Objectif

L'objectif est de désambiguïser un nom au sein d'un groupe de noms. Ce ne sont pas des synonymes, du moins pas nécessairement mais l'objectif est assez proche de Inkpen et Hirst. En revanche, la méthode semble plus convaincante[23].

Les groupes de noms sont de plus en plus utilisés pour des outils de TAL, notamment pour des thésaurus, en tant que classes sémantiques ou simple groupes de mots. Le fait est que ces outils auraient plutôt besoin de groupes de sens que de groupes de mots. Par exemple, lorsque l'on construit une classe de noms regroupant des professions de santé telles que *doctor* et *nurse*, on parle bien de *doctor* en tant que médecin et non en tant que détenteur d'un doctorat. De la même manière, on parle de *nurse* en tant qu'infirmière et non en tant que nourrice. Resnik propose de construire non des groupes de noms, mais des groupes de sens.

Le modèle prend en entrée un mot polysémique appartenant à un groupe de mots. Et il propose en sortie le ou les sens du mot les plus plausibles dans ce groupe. Plus précisément, il calcule un degré de plausibilité pour chacun des sens possibles du mot. Les sens dont nous parlons sont dans ces travaux les synsets de WordNet.

Méthode

Le cœur de l'algorithme de désambiguïsation est un calcul de similarités sémantiques entre sens de mots en utilisant la taxonomie de WordNet. L'intuition sous-jacente est que lorsque plusieurs mots cooccurrent, le sens le plus probable de chacun de ces mots est celui qui maximise les relations sémantiques entre chaque sens choisi. C'est-à-dire que le sens le plus probable d'un mot polysémique sera celui qui permet de réduire au maximum le nombre de liens taxonomiques avec les autres mots du groupe. Par exemple, si l'on reprend le cas de *doctor* et *nurse*, les sens 'médecin' et 'infirmière'

[23] On trouvera une approche analogue chez Sussna (1993).

sont très proches dans la taxonomie puisqu'ils sont reliés par l'hyperonyme « professionnel de la santé », alors que les sens « détenteur d'un doctorat » et « infirmière », ou encore « détenteur d'un doctorat » et « nourrice » sont beaucoup plus éloignés dans la taxonomie puisque que cela nécessiterait de remonter jusqu'à l'hyperonyme « individu », c'est-à-dire très haut dans la hiérarchie conceptuelle. Là où cette méthode se distingue, c'est que Resnik ne se contente pas de rechercher le plus court chemin entre deux sens, mais il calcule une distance plus élaborée, basée sur la distribution des mots étudiés dans un corpus de référence.

Calcul des similarités sémantiques

Comme nous l'avons dit, ce calcul est basé sur la taxonomie de WordNet, plus précisément sur les relations *IS-A* entre noms. La similarité sémantique entre deux mots w_1 et w_2 est calculée comme suit :

$$sim(w_1, w_2) = \max_{c \in subsumers(w_1, w_2)} [-\log \Pr(c)] \qquad (1)$$

où, *c* représente un concept, c'est-à-dire dans ce cas un synset de WordNet.

subsumers(w_1,w_2) est l'ensemble des concepts qui sont ancêtres de w_1 et w_2 dans l'un des sens de ces mots.

Pr(*c*) représente la probabilité d'apparition de *c* dans le corpus de référence. Cette probabilité correspond à la fréquence relative :

$$\Pr(c) = \frac{freq(c)}{N} \qquad (2)$$

où *N* est le nombre total d'instances de noms observé dans le corpus,

et $freq(c) = \sum_{n \in words(c)} count(n)$ (3)

où *words(c)* est l'ensemble des noms qui ont un sens dont l'un des ancêtres est le concept c.

Ainsi, le concept *c* qui maximise l'expression (1) est celui, parmi l'ensemble des *subsumers*, qui est le plus informatif pour w_1 et w_2. C'est-à-dire que c_1 *IS-A* c_2 implique $Pr(c_2) >= Pr(c_1)$. Cela garantit que le plus abstrait implique le moins informatif.

Ce calcul permet pour chaque paire de noms d'obtenir un degré de similarité sémantique et l'ancêtre le plus informatif pour cette paire. La figure 3.4 présente un extrait de ce que donne le calcul à partir de la version *Penn Treebank* du corpus Brown.

Word 1	Word 2	Similarity	Most Informative Subsumer
doctor	nurse	9.4823	⟨health professional⟩
doctor	lawyer	7.2240	⟨professional person⟩
doctor	man	2.9683	⟨person, individual⟩
doctor	medicine	1.0105	⟨entity⟩
doctor	hospital	1.0105	⟨entity⟩
doctor	health	0.0	*virtual root*
doctor	sickness	0.0	*virtual root*

Figure 3.4 : Extrait de résultat du calcul de similarité

Cet extrait montre les degrés de similarité que l'on peut calculer et notamment le fait que le nom *lawyer* (juriste/avocat) est plus éloigné de *doctor* que *nurse*, le concept qui relie *doctor* et *lawyer* étant « *professional person* » , un concept plus général que « *health professional* » qui réunit *doctor* et *nurse*. Cela semble correspondre à la mesure de similarité sémantique attendue. La présence du concept « *virtual root* », notamment entre *doctor* et *health* s'explique par le fait qu'il a fallu remonter jusqu'au niveau le plus haut de la hiérarchie, c'est-à-dire la racine virtuelle de l'ensemble des concepts, pour relier ces deux noms. De ce fait, le degré de similarité entre ces deux noms est nul.

Algorithme de désambiguïsation

L'algorithme est basé sur l'idée que lorsque deux mots polysémiques sont similaires/proches, leur ancêtre le plus informatif fournit des informations sur le sens à choisir pour chaque mot.

Etant donnés deux mots, w_1 et w_2 possédant différents sens, on attribue à chaque sens contenant l'ancêtre le plus informatif de la paire (w_1,w_2) parmi ses ancêtres, le degré de similarité calculée s(w_1,w_2). Pour un groupe de mots donnés, ce processus est

effectué pour chaque paire de mots possible. A la fin du processus, les sens d'un mot qui ont le plus de liens avec les autres mots du groupe sont ceux dont les degrés cumulés de similarité donneront le total le plus élevé.

Résultats

Resnik illustre ses résultats en présentant sept groupes de noms. Pour chaque groupe, il donne la source du groupe de noms, le groupe de noms, et pour chaque nom la liste de ses sens avec la valeur calculée par l'algorithme. La figure 3.5 présente l'un de ces groupes de noms, ainsi que les notes calculées pour chaque synset de chaque nom.

Distributional neighborhood (Schütze, 1993): burglars, thief, rob, mugging, stray, robbing, lookout, chase, crate

Word 'burglars' (1 alternatives)
1.0000 burglar: subconcept of thief, robber

Word 'thief' (1 alternatives)
1.0000 thief, robber: subconcept of criminal, felon, crook, outlaw

Word 'mugging' (1 alternatives)
1.0000 battering, beating, mugging, whipping: subconcept of fight, fighting

Word 'stray' (1 alternatives)
1.0000 alley cat, stray: homeless cat

Word 'lookout' (4 alternatives)
0.6463 lookout, lookout man, sentinel, sentry, watch, scout
0.0000 lookout, observation post: an elevated post affording a wide view
0.1269 lookout, observation tower, lookout station, observatory:
0.2268 lookout, outlook: subconcept of look, looking at

Word 'chase' (1 alternatives)
1.0000 pursuit, chase, follow, following: the act of pursuing

Word 'crate' (2 alternatives)
0.0000 crate, crateful: subconcept of containerful
1.0000 crate: a rugged box (usually made of wood); used for shipping

Figure 3.5 : Traitement d'un groupe de noms

On peut noter que pour le mot polysémique *lookout*, l'algorithme a permis de privilégier le sens de 'guetteur', 'sentinelle', 'garde' plutôt que les autres sens qui font référence à des lieux (poste d'observation, tour d'observation, observatoire) ou encore à une action (regarder attentivement). Pour le deuxième mot polysémique *crate*, c'est le sens de caisse en tant qu'objet qui est privilégié par rapport au sens d'unité de mesure.

Resnik présente ensuite une évaluation de ses résultats à l'aide de deux juges : le juge 1 doit tester 99 noms polysémiques. Pour chaque nom, on présente au juge le groupe de noms dans lequel il se trouve. Le juge doit choisir un et un seul sens parmi les sens du nom proposés par WordNet. Il doit ensuite ajouter une note de fiabilité à son choix : 0 si ce choix est peu fiable et 4 s'il est très fiable. Comme point de référence ou borne inférieure (*baseline*), dix noms ont été testés en sélectionnant un sens au hasard, ce qui a donné 34,8% de sens corrects. Comme borne supérieure[24], le juge 2 était correct sur 65,7% des instances à tester. L'algorithme de désambiguïsation est proche de cette borne supérieure avec 58,6% de sens corrects.

Resnik fait ensuite la même opération en inversant le rôle des juges et obtient des résultats similaires.

Discussion

Ce travail illustre une manière d'optimiser l'utilisation d'un dictionnaire comme source de connaissance. Le point crucial étant de proposer un calcul de similarité sémantique qui n'est pas simplement basé sur la structure du dictionnaire, mais aussi sur le comportement des mots dans un corpus de référence. En cela on peut classer cette étude parmi les approches mixtes.

Un autre point important est que l'objectif n'est pas de donner un et un seul sens correct pour un mot dans un groupe de mots, mais plutôt de classer les sens de ce mot, du plus probable au plus improbable.

Selon les propres termes de l'auteur, ce n'est cependant pas un modèle de désambiguïsation à proprement parler, mais plutôt un module à intégrer dans un modèle plus global qui prendrait davantage en considération les usages contextuels des noms.

[24] La borne considérée par l'auteur comme l'efficacité maximum que l'on puisse atteindre. Nous développerons ce point dans le paragraphe 3.2.2.

Remarque : les mots dont le degré de fiabilité donné par le juge était égal à zéro ou un (c'est-à-dire tous les jugements peu fiables) ne sont pas pris en compte dans les résultats de l'évaluation. C'est peut-être une bonne chose, néanmoins il serait intéressant de voir si ces mots exclus ne sont pas justement les plus difficiles à désambiguïser.

3.1.6 Schütze (1998)

Objectif

Nous terminons notre tour d'horizon des méthodes de désambiguïsation par l'étude qui est sans aucun doute la plus proche de nous. Schütze propose de diviser en deux la tâche de désambiguïsation d'un mot : la discrimination des sens et la dénomination des sens. La discrimination de sens divise les occurrences d'un mot en un certain nombre de classes en déterminant pour chaque couple d'occurrences si elles renvoient au même sens ou non. La dénomination du sens donne une étiquette sémantique à chaque classe, et à l'aide de la discrimination de sens, à chaque occurrence du mot ambigu.

L'auteur considère que la tâche de discrimination est plus facile qu'une désambiguïsation complète puisqu'on a juste à déterminer quelles sont les occurrences qui ont le même sens et pas à dire quel est ce sens. C'est ce qu'il propose de faire dans cette étude. L'intérêt ici étant de se passer d'une ressource externe pour nommer les sens. Nous avons donc affaire à une approche purement endogène. Ce travail est suffisant pour des tâches de recherche d'information où l'on n'a pas besoin de nommer les sens, juste de rassembler ceux qui sont proches.

Méthode

Le but de cette méthode est de grouper les occurrences d'un mot ambigu en clusters, où les clusters sont construits à partir de similarités contextuelles des occurrences.

Pour cela, mots, contextes et clusters sont représentés dans un espace vectoriel à grande dimension, muni de la distance utilisée dans l'analyse sémantique latente (LSA : *Latent Semantic Analysis*, cf. Deerwester *et al.* 1990)[25]. Les vecteurs de contextes capturent l'information présente dans les co-occurrences de second ordre, c'est-à-dire les cooccurrences des cooccurrences.

Les vecteurs de contextes sont regroupés de telle manière que les occurrences qui sont jugées similaires d'après leurs cooccurrences du second ordre sont mises dans le même cluster. Les clusters sont représentés par leur centre de gravité. Une occurrence dans un texte test est désambiguïsée en calculant la représentation de second ordre du contexte étudié, et en l'assignant au cluster dont le centre de gravité est le plus proche de la représentation.

Représentation vectorielle

La méthode consiste d'abord à représenter vectoriellement les mots w_i du contexte de l'occurrence du mot w à désambiguïser. A chaque w_i on fait correspondre un vecteur V_i dont les coordonnées sont obtenues en considérant les contextes de toutes les occurrences du mot w_i. Le contexte d'une occurrence est défini par une fenêtre de cinquante mots autour de l'occurrence en question. Deux variantes ont été étudiées par Schütze : une méthode dite globale, où l'espace vectoriel de représentation est défini une fois pour toutes, indépendamment du mot à désambiguïser, et une méthode dite locale, où un espace vectoriel spécifique est dédié à chaque mot à désambiguïser.

Dans la méthode globale, les dimensions de l'espace vectoriel sont constituées par les 2 000 mots[26] les plus fréquents dans le corpus. On calcule alors, une fois pour toutes, les coordonnées des vecteurs représentant les 20 000 mots les plus fréquents

[25] On notera aussi les travaux de Straparava, Gliozzo et Giuliano (2004) qui exploitent les vecteurs LSA parmi d'autres sources de connaissances dans un modèle combinant apprentissage supervisé et non-supervisé.

[26] Il s'agit de mots pleins : une liste de mots-outils (*stop list*) est exclue en premier lieu.

dans cet espace : la coordonnée c_{ij} du mot w_i dans la dimension w_j est égale au nombre de fois que w_j se trouve dans le contexte de w_i. Une occurrence du mot w à désambiguïser est alors représentée dans le même espace par la moyenne des vecteurs associés aux mots w_i présents dans ce contexte particulier de w.

Dans la méthode locale, le principe est le même, sauf que pour un mot donné w à désambiguïser, les dimensions de l'espace vectoriel sont constituées par 1 000 mots w_i sélectionnés dans les contextes des occurrences de w, et ce sont ces mêmes w_i dont on calcule une représentation vectorielle dans cet espace. La sélection de ces w_i peut s'effectuer par plusieurs méthodes. Celle qui donne les meilleurs résultats est une méthode qui utilise une mesure du χ^2 : on choisit les mots dont la présence est le mieux corrélée à la présence de w selon cette mesure.

Clusterisation et désambiguïsation

La discrimination des sens d'un mot ambigu est opérée par la construction de clusters de vecteurs représentant les différentes occurrences de ce mot dans le corpus d'apprentissage. La distance utilisée est la distance classique de la LSA : on opère d'abord une réduction de la dimensionnalité de l'espace par une analyse en composantes principales, puis on utilise le cosinus de l'angle entre deux vecteurs dans cet espace réduit comme distance entre les deux occurrences correspondantes.

Pour chaque cluster obtenu, on calcule son barycentre et on lui associe un des sens du mot ambigu. Pour désambiguïser une nouvelle occurrence du mot, il suffit alors de calculer la distance du vecteur associé à cette occurrence à chacun des barycentres des clusters et de choisir le sens correspondant au barycentre le plus proche.

Résultats

Schütze a utilisé comme corpus d'apprentissage et de test des dépêches du *New York Times*. L'apprentissage a été effectuée sur un corpus de grande taille (plus de 60 millions de mots) et le test sur un autre corpus (de 5 millions de mots) issu de la même source.

Pour l'évaluation, deux types de test ont été réalisés. D'une part, dix mots ambigus ont été choisis, avec deux sens très différents (homonymes), tels que *suit* (costume ou poursuite judiciaire) et *plant* (usine ou plante). Pour ces mots, c'est manuellement qu'un des deux sens a été attribué aux différentes occurrences dans le corpus de test pour les comparer aux sorties du système. D'autre part, dix ambiguïtés « artificielles » ont été construites : deux mots très différents, comme *banane* et *fenêtre*, ont été fusionnés en un même « mot », conduisant ainsi artificiellement à une ambiguïté. Pour ces faux mots fusionnés, on évite l'étiquetage manuel des sens puisque l'on sait pour chaque occurrence quel mot réel était présent au départ.

Les résultats montrent que sur l'ensemble c'est la méthode globale qui est la meilleure (de l'ordre de 90% de succès), mais il est intéressant de constater que si l'on se restreint aux ambiguïtés « naturelles », la méthode locale (utilisant la sélection par la mesure du χ^2) est aussi bonne, voire meilleure : c'est surtout sur les ambiguïtés artificielles que la méthode globale réussit le mieux.

Discussion

Ces très bonnes performances ne sont pas particulièrement impressionnantes puisque, comme on l'a déjà fait remarquer, pour les tâches de désambiguïsation grossières entre sens homonymiques, de tels taux de réussite sont assez courants. Néanmoins, cette étude est très intéressante à bien des égards. D'abord, il s'agit d'une méthode entièrement automatique, même si elle se limite à la discrimination des sens, laissant de côté le problème de leur dénomination. Mais c'est surtout au plan théorique que son apport est important pour nous. En effet, la représentation géométrique qui est à la base du modèle permet d'associer à chaque occurrence un vecteur dans un espace, la proximité de deux vecteurs étant censée refléter la proximité des sens correspondants du mot étudié. Comme on le verra, notre modèle, s'il diffère par bien des aspects, est lui aussi fondé sur des représentations géométriques du même genre. De plus, au plan technique, nous utilisons aussi en grande partie les mêmes outils (réduction de dimensionnalité et mesure du χ^2 notamment).

3.2. Evaluation et confrontation des modèles

Parmi les travaux présentés, nous avons parfois émis quelques réserves concernant des points de méthode ou des évaluations abusives. Notamment, sur le choix de mots annoncés comme polysémiques mais qui relèvent plus de l'homonymie comme dans le travail de Peng *et al.* (2001), ou encore sur l'utilisation abusive de la notion de « relation syntaxique » lorsque les auteurs n'emploient que des patterns de collocation simples.

Dans cette partie, nous voulons mettre en évidence certaines questions récurrentes dans le domaine de la désambiguïsation concernant l'évaluation et la confrontation de ces modèles. Ces questions découlent de plusieurs points de divergence. Tout d'abord, les quelques études que nous avons présentées nous montrent que les objectifs divergent. Certains cherchent à désambiguïser un nombre restreint de mots alors que d'autres veulent pouvoir désambiguïser tous les mots d'un texte. Certains s'intéressent au sens d'un mot dans un groupe de mots alors que d'autres s'intéressent au sens d'un mot dans un texte. Le deuxième point de distinction concerne le degré de désambiguïsation, c'est-à-dire la granularité du sens recherché. Certains modèles en restent au niveau homonymique alors que d'autre cherchent une désambiguïsation beaucoup plus fine. Enfin le dernier point concerne l'évaluation. Comment juger si un modèle est efficace ou non ? Qu'est ce qu'un modèle de désambiguïsation mauvais, et qu'est ce qu'un modèle de désambiguïsation parfait ?

Nous commencerons par développer ces trois points. Nous allons ensuite décrire la campagne Senseval dont l'objectif est justement de créer un *benchmark* commun afin de contrôler ces divergences et de confronter sur les mêmes bases des modèles différents. Nous terminerons par une discussion sur les nouveaux problèmes posés par Senseval.

3.2.1 Que peut-on comparer ?

Pour commencer, il y a les études qui ne proposent pas le même type de résultats. Le cas le plus net est celui de Schütze (1998) qui distingue deux étapes dans la désambiguïsation : la discrimination de sens et la dénomination de sens. Dans son étude, il ne s'intéresse qu'à la première étape, par conséquent les résultats qu'il cherche à obtenir ne sont pas des sens de mots mais des rapprochements de mots. Il semble alors bien difficile de comparer des distances entre mots avec les résultats obtenus par un modèle plus classique proposant un ou plusieurs items de sens pour désambiguïser un mot.

Si l'on reste dans cette dernière catégorie de travaux, la comparaison reste encore difficile. Les modèles de Resnik (1995) et Inkpen et Hirst (2003) cherchent à désambiguïser le sens d'un mot dans un groupe de mots et non dans un texte comme la plupart du temps. Inkpen et Hirst admettent eux-mêmes que leur tâches est beaucoup plus facile qu'une désambiguïsation en contexte puisque leurs groupes de mots sont extrait d'un dictionnaire de synonymes contenant pour chaque entrée un texte descriptif ainsi que différentes sources d'informations (hyperonymes, antonymes, etc.). Ces deux modèles ne sont pas comparables pour autant puisque les groupes de mots que Resnik cherche à désambiguïser sont vierges de toute autre information.

Regardons maintenant la tâche précise « désambiguïsation du sens d'un mot dans un texte par l'attribution d'un ou de plusieurs sens d'un dictionnaire ». Ici encore, il faut distinguer deux types d'approches qu'il serait sans intérêt de comparer. D'un côté, il y a ceux qui cherchent à désambiguïser un petit groupe de mots et de l'autre ceux qui cherchent à désambiguïser n'importe quel mot d'un texte. L'incidence est importante tant sur le modèle que sur les résultats. En effet, les premiers vont baser le modèle sur la liste de mots à désambiguïser alors que les seconds vont construire un modèle basé sur des règles plus générales applicables à tous les mots. Ainsi, les résultats seront nettement meilleurs pour les premiers au prix d'un modèle spécifique à une liste de mots.

3.2.2 La granularité de sens

Nous avons une illustration caricaturale de ce phénomène avec la comparaison des travaux de Wilks et Stevenson (1998) d'une part, et ceux de Peng, Ito et Furugori (2001) d'autre part. Ils ont un objectif commun qui est de « désambiguïser le sens de n'importe quel mot dans un texte par l'attribution d'un ou de plusieurs sens d'un dictionnaire ». La précision annoncée par Peng, Ito et Furugori (2001) pour cette tâche est de 92,1%. Dans le modèle de Wilks et Stevenson, le résultat de la désambiguïsation est une liste ordonnée de sens. Ils annoncent un taux de réussite de 70% pour que le sens étiqueté soit le premier de la liste. Les résultats ne sont pas pour autant comparables, et il ne fait aucun doute selon nous que les résultats obtenus par Wilks et Stevenson sont bien plus concluants. En effet, le taux de réussite annoncé par Peng, Ito et Furugori correspond à un choix entre deux sens homonymiques (*sentence* : group of word/punishment ; *crane* : machine/bird ; *plant* : living thing/factory ; etc.) alors que Wilks et Stevenson s'intéressent aux sens les plus fins du LDOCE, avec un taux moyenne de polysémie de 7,65[27].

3.2.3 Bornes inférieures et bornes supérieures

Il s'agit ici de replacer les taux de précision obtenus par un modèle dans un intervalle raisonnable. La borne inférieure correspond à la valeur en dessous de laquelle le modèle n'est pas intéressant. La borne inférieure la plus fréquemment utilisée correspond à la précision obtenue si l'on attribue à chaque instance d'un mot, son sens le plus fréquent. Cette précision est toujours supérieure à la précision obtenue s'il l'on attribue à chaque instance d'un mot, un sens aléatoire. En effet, il est très rare que la distribution des sens d'un mot dans un corpus soit équiprobable. Prenons le cas du nom *chef.* Audibert (2003) en décrit onze sens différents et montre que si l'on attribue à chaque occurrence de *chef* un des onze sens au hasard, la précision est de 9%. Si l'on

[27] Cf. paragraphe 5.1 ainsi que Veronis (1998, 2004) pour un prolongement de la question de la granularité.

attribue à chaque occurrence le sens le plus fréquent, la précision, appelée précision majoritaire, est de 76%. Dans ce cas la précision majoritaire sera probablement difficile à dépasser. Plus généralement, cette précision majoritaire est aux alentours de 50%, comme par exemple dans l'étude d'Escudero *et al.* (2000) où la précision majoritaire pour les noms est de 56,4% et pour les verbes de 46,7%. Par comparaison, la précision au hasard dans l'étude de Resnik est de 34,8%.

La borne supérieure dépend beaucoup plus de la méthode d'évaluation utilisée dans l'étude. Le principe général est de dire qu'un modèle de désambiguïsation automatique ne peut pas faire mieux qu'un humain. On prend alors les résultats donnés par un humain comme référence, et on confronte le modèle automatique à cette référence. Se pose alors la question de *l'accord interjuge,* c'est à dire, quelle est la précision des résultats fournis par un humain h1 si l'on prend comme référence les résultats donnés par un humain h2. Le désaccord est particulièrement important dans l'étude de Resnik. Les choix du juge 1 on une précision de 65,7% par rapport à ceux de juge 2. Dans les modèles basés sur un apprentissage supervisé, tel celui de Yarowsky (2000), la question ne se pose pas au moment de la désambiguïsation, puisque l'on utilise un corpus annoté sémantiquement qui sert de référence. Le problème est simplement déplacé puisque l'accord interjuge intervient au moment de l'annotation du corpus. Véronis (1998, 2004) décrit ce problème et la grande variabilité de l'accord interjuge en fonction de la granularité de sens désirée, du dictionnaire utilisé, et de la manière de regrouper les sens.

3.2.4 Senseval

Kilgarriff part du constat de Yarowsky. Ce dernier évoquait en 1992 la difficulté de comparer les résultats entre expériences sur la désambiguïsation, étant donné leurs différences importantes, notamment au niveau des corpus, des mots, des juges, des calculs de précision, et des outils utilisés tels que les étiqueteurs grammaticaux.

Kilgarriff propose alors en 1998 de construire un standard à partir duquel tous les modèles puissent se mesurer. Ce sera la première campagne de Senseval. L'objectif est de savoir quel est le meilleur modèle de désambiguïsation automatique, quels sont les mots et les langues qui posent des problèmes particuliers à quels modèles. Lors de cette première campagne, il n'y a qu'une tâche évaluée, la désambiguïsation d'un nombre limité de mots (« lexical sample task »), sur trois langues, l'anglais, le français et l'italien. Cette évaluation est faite sur 15 noms, 13 verbes, 8 adjectifs et 5 adverbes.

Pour la troisième édition de Senseval, qui s'est terminée en 2004, le nombre de tâches a augmenté, ainsi que le nombre de langues[28]. On retrouve les tâches « *all words* » et « *lexical sample* », qui correspondent respectivement aux modèles capables de désambiguïser automatiquement n'importe quel mot du texte et ceux adaptés à une liste limitée de mots. Les données utilisées pour la tâche « *lexical sample* » sont des exemples extraits du BNC (British National Corpus) et annotés sémantiquement à l'aide de WordNet 1.7.1. Il y a 60 mots à désambiguïser (noms, adjectifs, verbes ambigus). Enfin, cette campagne propose deux sous évaluations pour la tâche « *lexical sample* » : calcul de sens grossiers et calcul de sens fins.

La participation est très importante : rien que pour l'anglais, la tâche « *all word* » a mobilisé 64 équipes, et la tâche « *lexical sample* » en a mobilisé 65. Cette dernière tâche avait mobilisé 17 systèmes lors de la première campagne en 1998. Cette troisième campagne a confirmé le fait que les meilleurs résultats obtenus pour la tâche « *lexical sample* » et pour la tâche « *all words* » proviennent de modèles basés sur apprentissage supervisé. La nouveauté étant que les modèles qui se démarquent sont ceux qui comme Straparava (2004) introduisent des critères non supervisés tels que les vecteurs LSA.

[28] Liste des tâches de la troisième campagne Senseval : English all words ; Italian all words ; Basque lexical sample ; Catalan lexical sample ; Chinese lexical sample ; English lexical sample ; Italian lexical sample ; Romanian lexical sample ; Spanish lexical sample ; Automatic subcategorization acquisition ; Multilingual lexical sample ; WSD of WordNet glosses ; Semantic Roles ; Logic Forms.

Il est important de noter que globalement ces campagnes d'évaluation répondent aux problèmes que nous avons évoqués précédemment : les différences de granularité de sens, les problèmes de bornes supérieures sont évités en comparant les modèles entre eux et enfin la distinction entre des tâches « *all words* » et des tâches « *lexical sample* ».

Remarques

On ne peut que louer les mérites d'une telle campagne, et le nombre croissant d'équipes qui s'inscrivent à chaque campagne montre l'intérêt encore grandissant de la désambiguïsation automatique. De plus, le fait que l'ensemble des épreuves soit basé sur des annotations sémantiques provenant de WordNet aurait pu être critiquable. En effet, on a longtemps reproché à WordNet d'être trop parcellaire. C'est-à-dire que mis à part les synsets et les relations *IS-A*, le reste était quasi inexistant. Pas de liens entre les parties du discours (noms, verbes, adjectifs), pas de connaissances encyclopédiques, pas d'indication de domaine, pas d'information sur les propriétés syntaxiques, etc. Il y a eu un énorme effort dans ce domaine, et actuellement il est difficile de critiquer cette énorme avancée et le résultat obtenu. Prenons le cas du mot *mean* et son entrée dans WordNet (cf. Figure 3.6. Nous n'avons pas reporté le sens adjectival de *mean : the mean annual rainfall, a mean person, etc.* Pour l'intégralité de la définition : http://wordnet.princeton.edu/perl/webwn). On peut noter que pour le sens nominal de *mean*, WordNet propose un domaine qui est ici la statistique. Concernant le synset *« mean, intend »*, WordNet propose des formes grammaticales dérivées, le mot du synset avec lequel elles sont reliées, et des exemples d'emploi. WordNet propose aussi des structures de phrase : « somebody …s something » et « somebody …s somebody ».

La deuxième remarque est que si la langue française faisait partie de la première campagne Senseval de 1998, ça n'est plus le cas aujourd'hui. C'est très regrettable puisqu'il n'existe pas de *benchmark* équivalent en France, et les modèles de désambiguïsation basés sur le français se trouvent dans la situation d'avant Senseval : ils ne peuvent pas se comparer objectivement. Peut-être faut-il profiter de cette lacune

pour proposer un autre mode d'évaluation. Dans le cas de l'anglais, WordNet prend une telle ampleur que du statut de simple ressource lexicale il devient une référence pour tous, ce qui implique un formatage « WordNetien » des modèles de désambiguïsation anglais. Quelle que soit la qualité atteinte par WordNet, il n'est pas raisonnable de baser tout un secteur de recherche sur une seule ressource lexicale.

La solution pourrait être une évaluation basée sur des applications, comme l'a déjà proposé Schütze (1998). C'est-à-dire comparer les modèles de désambiguïsation sur une tâche appliquée bien précise telle que l'indexation de documents, la traduction ou encore la recherche d'information. Il est probable que ce type d'évaluation soit tout aussi complexe à mettre en place. Nous en resterons à ces suggestions dans cette étude mais il semble que ce soit une solution d'évaluation à développer.

1.1.1 Noun

- S: (n) mean, mean value (an average of n numbers computed by adding some function of the numbers and dividing by some function of n)
    - *direct hyponym* / *full hyponym*
    - *domain category*
        - S: (n) statistics (a branch of applied mathematics concerned with the collection and interpretation of quantitative data and the use of probability theory to estimate population parameters)
    - *direct hypernym* / *inherited hypernym* / *sister term*

1.1.2 Verb

- S: (v) mean, intend (mean or intend to express or convey) *"You never understand what I mean!"; "what do his words intend?"*
    - *direct troponym* / *full troponym*
    - *direct hypernym* / *inherited hypernym* / *sister term*
    - *derivationally related form*
        - W: (n) meaning [Related to: mean] (the message that is intended or expressed or signified) *"what is the meaning of this sentence"; "the significance of a red traffic light"; "the signification of Chinese characters"; "the import of his announcement was ambiguous"*
        - W: (n) meaning [Related to: mean] (the idea that is intended) *"What is the meaning of this proverb?"*
        - W: (n) intent [Related to: intend] (the intended meaning of a communication)
    - ***sentence frame***
        - Somebody ----s something
        - Somebody ----s somebody
- S: (v) entail, imply, mean (have as a logical consequence) *"The water shortage means that we have to stop taking long showers"*
- S: (v) mean, intend, signify, stand for (denote or connote) *"`maison' means `house' in French"; "An example sentence would show what this word means"*
- S: (v) intend, mean, think (have in mind as a purpose) *"I mean no harm"; "I only meant to help you"; "She didn't think to harm me"; "We thought to return early that night"*
- S: (v) mean (have a specified degree of importance) *"My ex-husband means nothing to me"; "Happiness means everything"*
- S: (v) think of, have in mind, mean (intend to refer to) *"I'm thinking of good food when I talk about France"; "Yes, I meant you when I complained about people who gossip!"*
- S: (v) mean (destine or designate for a certain purpose) *"These flowers were meant for you"*

Figure 3.6 : Exemple d'entrée de la ressource lexicale WordNet

3.3. Deux sources de connaissances essentielles : les relations syntaxiques et les classes sémantiques

Nous l'avons déjà répété plusieurs fois, les modèles de désambiguïsation varient énormément d'une équipe à l'autre. Les distinctions concernent les méthodes employées (utilisation de réseaux sémantiques, construction de vecteurs contextuels ou encore apprentissage supervisé), les ressources utilisées (dictionnaires, corpus, Web), enfin les sources de connaissances exploitées (collocations, relations pragmatiques, relations syntaxiques, etc.).

Ce que l'on peut constater, au vu des travaux présentés ici ainsi que des résultats obtenus lors de la troisième campagne d'évaluation de Senseval, c'est que la mixité prévaut. C'est-à-dire que les modèles les plus efficaces sont ceux qui exploitent plusieurs ressources (corpus, dictionnaires, ontologies) et plusieurs sources de connaissances (Stevenson et Wilks 2001, Yarowsky 2000, Straparava 2004). Agirre et Martinez (2001) proposent de classer les différents systèmes en fonction des ressources qu'ils utilisent. Ils énumèrent aussi toutes les sources de connaissances qui peuvent être utilisées. En voici la liste :

Les parties du discours (en anglais POS pour Part Of Speech) sont utilisées pour organiser les sens de mots. Par exemple, distinguer dans un corpus les sens de *avoir* en tant que verbe (*avoir vingt ans*) et en tant que nom (*donner un avoir à quelqu'un*).

La morphologie, spécialement les relations entre mots dérivés d'une même forme morphologique. Par exemple les relations entre le nom *beauté* et l'adjectif *beau*. En tenant compte du fait qu'il n'y a pas une superposition totale entre les deux emplois. Une belle fortune ne parle pas de la beauté de la fortune mais de sont importance quantitative.

Les cooccurrences : Deux types d'informations véhiculées par les cooccurrences semblent se distinguer : les informations locales et les

informations globales. Les informations locales correspondent à des dépendances locales comme l'adjacence au mot à désambiguïser, l'appartenance à un contexte réduit contenant ce mot. Les informations globales consistent en de larges fenêtres contenant 50 à 100 mots sous la forme de lemmes.

Les associations sémantiques de mots

Organisation en taxonomie par exemple associer *table* à *meuble* par une relation hyperonymique.

Associations thématiques, comme entre *batte* et *baseball*

Les indices syntaxiques : par exemple, le sens *prendre un repas* du verbe *manger* est intransitif tandis que ses autres sens (avaler pour se nourrir, ronger, absorber, consommer, etc.) sont transitifs.

Les rôles actantiels le fait de savoir que le sujet du verbe *jouer* est un individu et non un objet permet de contraindre les sens possibles pour *jouer*.

Les restrictions de sélection permettent, par exemple, de préciser que le verbe *manger*, employé dans le sens prendre un repas, préfère un sujet de type humain.

Le domaine. Par exemple, dans un domaine sportif, on préférera le sens « raquette de tennis » pour le mot *raquette*.

La fréquence des sens. La distribution des sens d'un mot polysémique et très rarement équiprobable dans un corpus. Un seul sens peut parfois rassembler 90% des instances observées. Cet indicateur peut alors s'avérer utile, notamment dans le cas où aucune autre source d'information ne permet de trancher.

Les équipes travaillant dans ce domaine s'accorderaient pour dire que toutes ces informations sont potentiellement utiles pour un modèle de désambiguïsation. Agirre et Martinez (2001) ont montré que les relations de cooccurrences semblaient être la source d'information la plus efficace pour la désambiguïsation lexicale. Cela n'est pas surprenant à partir du moment où la grande majorité des études sont faites sur la désambiguïsation des noms. Si l'on s'intéresse aux verbes, les priorités changent. Il est clair que l'étude des cooccurrences lexicales jouera encore un rôle important, mais ne sera plus suffisante. On peut d'ailleurs noter qu'à l'exception de l'étude de Inkpen et Hirst (2003), l'ensemble des modèles est moins efficace pour la désambiguïsation des verbes que pour celle des noms.

Cela s'explique par le fait que le sens d'un verbe dépend plus particulièrement de deux sources : d'une part de la construction syntaxique dans laquelle il est employé et d'autre part de la classe sémantique des éléments lexicaux qui lui sont rattachés (cf. § 1.2). Ce sont ces deux sources de connaissance qui prendront une place centrale dans notre étude. Or ces sources posent des problèmes importants pour le traitement automatique.

Un nombre important de modèles revendique la prise en compte de relations syntaxiques dans leur modèle de désambiguïsation. Cette notion de relation syntaxique dissimule un panel d'indices allant de la simple position relative des mots à des relations syntaxiques telles que la relation verbe-objet indirect. Parmi les premiers modèles de désambiguïsation automatique revendiquant la prise en compte de l'influence syntaxique on peut noter ceux de Yarowsky (1992) et Ng et Lee (1996) qui utilisent des propriétés telles que les dépendances locales autour du mot étudié : collocations, relations de tête d'argument et des formes syntaxiques limitées telles que deux POS (Part Of Speech) adjacents pour Yarowsky. Ng et Lee (1996) incluent la notion de relation verbe-objet en considérant que « si un nom à désambiguïser [en anglais] est la tête d'un groupe nominal, indiqué par sa position à la fin du groupe nominal repéré, et si le mot qui précède immédiatement ce groupe nominal est un verbe,

ils considèrent que le couple verbe-nom est dans une relation syntaxique verbe-objet. ». Autrement dit, le terme de relations syntaxiques est un peu excessif dans ces deux études et les relations présentées ne sont exploitées que pour les noms.

Yarowsky (2000) exploite un analyseur syntaxique, mais comme nous l'avons vu l'objectif n'est pas le même que le nôtre puisque Yarowsky développe un modèle pour la désambiguïsation d'un nombre limité de mots, nécessitant l'introduction de connaissances difficilement automatisables. De plus Yarowsky n'applique les relations syntaxiques que pour la désambiguïsation des noms.

A notre connaissance, les seules études exploitant réellement les relations syntaxiques pour la désambiguïsation des verbes sont celles de Wilks et Stevenson (1998), Stevenson et Wilks (2001). Ils utilisent l'analyseur syntaxique développé par Stevenson (1998), ce qui explique probablement les résultats particulièrement bons qu'ils annoncent.

Ceci étant, s'il exploitent les relations syntaxiques, ils n'exploitent pas directement l'information qu'elles contiennent. Elles servent de support à l'étude des restrictions de sélection. C'est-à-dire que l'énoncé dans lequel se trouve le mot à désambiguïser est transposé en construction actantielle par l'intermédiaire de la construction syntaxique de l'énoncé.

Nous avons déjà introduit la théorie des grammaires constructionnelles dont l'hypothèse centrale est de considérer que les constructions syntaxiques sont porteuses d'un sens intrinsèque indépendamment du lexique. Un des objectifs de notre étude est de tester cette hypothèse, et donc donner aux constructions syntaxiques l'appellation de source de connaissance indépendante (cf. chapitres 7 et 8).

Revenons maintenant sur les constructions actancielles décrites par Wilks et Stevenson pour l'étude des restrictions de sélection. Les restrictions de sélection sont extraites du dictionnaire LDOCE. Les définitions du LDOCE contiennent des restrictions de sélection simples pour chaque mot. Pour les noms, un ensemble de 35

classes sémantiques est utilisé (*humain, humain male, solide, plant*, etc.). Pour les verbes on a les classes du sujet, objet direct, et objet indirect. La manière d'extraire les restrictions de sélection des définitions n'est pas précisée par les auteurs. De plus, les classes du LDOCE ne sont pas hiérarchisées, ils utilisent donc une hiérarchie construite manuellement (Bruce et Guthrie, 1992).

L'utilisation de classes construites à partir d'un dictionnaire pose un certain nombre de questions. Quelle liste de classes choisir ? Ici la liste est limitée à 35 classes. Comment hiérarchiser ces classes ? Ici c'est fait manuellement. Enfin, comment traiter les mots polysémiques appartenant à différentes classes ?

Nous proposons dans notre étude d'utiliser des classes sémantiques construites à partir d'une analyse distributionnelle sur corpus. Les travaux de ce type sont nombreux et nous proposons d'en décrire une partie ainsi que notre approche dans le sous-chapitre 9.1. Retenons pour l'instant que le calcul automatique de nos classes repose sur deux principes :

1- Nous ne construisons pas des classes sémantiques à proprement parler, mais des classes distributionnelles composées de mots qui influent de la même manière sur le sens des mots voisins.

2- Nous ne construisons pas des classes générales de la langue, mais des classes qui dépendent du verbe que l'on veut désambiguïser : autrement dit, la classe d'un nom dépend du contexte d'emploi de ce nom.

4. Un modèle dynamique de calcul du sens

Si l'on reprend les distinctions entre travaux endogènes, exogènes et mixtes du chapitre précédent, on peut considérer que nous faisons partie des approches mixtes puisque nous utilisons deux ressources différentes : un dictionnaire des synonymes (données exogènes) et un corpus (données endogènes). Le dictionnaire des synonymes va être utilisé pour calculer l'espace sémantique de chaque unité lexicale et le corpus pour le calcul de désambiguïsation à proprement parler. On peut aussi considérer que nous nous plaçons parmi les approches qui exploitent différentes sources de connaissances puisque nous exploitons d'une part l'influence du co-texte lexical, d'autre part l'influence de la construction syntaxique et enfin l'influence des classes sémantiques du co-texte lexical. En revanche, notre étude se distingue des approches que nous avons présentées sur trois points fondamentaux.

Tout d'abord, nous posons comme postulat que la construction syntaxique dans laquelle se situe le verbe est un élément du co-texte comme toute autre unité lexicale. Certes, certains modèles tiennent compte de la position du co-texte et même de la relation syntaxique que ce co-texte entretient avec l'unité étudiée (Yarowsky, 2000 ; Wilks et Stevenson 1998). La différence est qu'il ne s'agit plus de dire que le sens s_1 d'un mot implique telle ou telle construction verbale comme le font les restrictions de sélection, mais de considérer que la construction syntaxique, et donc la construction verbale, est porteuse d'un sens intrinsèque, indépendamment des unités lexicales qui la composent, et que ce co-texte « syntaxique » doit être intégré dans le calcul du sens. Ce postulat s'appuie sur la théorie des grammaires constructionnelles et nous développerons ce point dans le chapitre 7.

Dans la plupart des évaluations de modèles de désambiguïsation, les mots testés sont des mots très fréquents. En effet, pour que le calcul de désambiguïsation soit efficace, il faut avoir des informations fiables sur le comportement du co-texte, c'est-à-dire avoir des fréquences de cooccurrence fiables. Ce type de calcul ne peut pas s'appliquer à des cooccurrences rares du corpus. Autrement dit, on peut désambiguïser le sens du verbe *compter* dans *compter les moutons*, mais pas dans *compter les ouailles* (*mouton, et plus particulièrement brebis* selon la définition du TLFi). Nous proposons un moyen de combler cette lacune. Il consiste à remplacer chaque co-texte par une classe de mots qui influent de la même manière sur le mot à désambiguïser ; il s'agit par exemple d'être capable de remplacer *ouaille* par *{ouaille ; mouton ; brebis}*. Cela permet ainsi de retomber sur des co-textes interprétables en terme de fréquence. Plusieurs modèles exploitent des classes sémantiques (Resnik 1997 ; Wilks et Stevenson 1998 ; Kohomban 2005). Le point crucial est que nos classes ne sont pas extraites d'un thésaurus ou d'une autre source exogène, mais elles sont extraites du corpus. Et, plus précisément notre méthode permet un calcul de classes dépendant du contexte. Ainsi, dans un même texte, un mot comme *boîte* n'aura pas la même classe sémantique dans le contexte *travailler dans une boîte {boîte ; société ; entreprise}* que dans le contexte *ranger dans une boîte {boîte ; caisse ; carton}*.

Enfin, le dernier point d'originalité est que cette étude n'est consacrée qu'aux verbes : la désambiguïsation automatique a commencé par s'intéresser aux noms. Progressivement, les verbes et les adjectifs sont introduits dans les systèmes d'évaluation, mais il n'y a pas de prise en compte particulière de ces deux parties du discours, c'est-à-dire que l'on ne fait qu'appliquer aux verbes et aux adjectifs un modèle adapté à la désambiguïsation des noms. On ne peut que constater la chute des taux de précision pour les verbes et les adjectifs par rapport aux noms (à l'exception inexpliquée de Inkpen et Hirst 2003). Ces écarts de précision peuvent aussi s'expliquer par la relative nouveauté de l'exploitation de sources de connaissances multiples, et notamment de l'exploitation des contraintes syntaxiques.

Les trois distinctions que nous venons de présenter concernent des points de méthode ou d'objet d'étude. On pourrait alors considérer que cette étude est un prolongement des travaux du même domaine. Ce n'est pas le cas. En effet l'élément crucial qui distingue ce travail des précédents est le fondement linguistique sous-jacent. Ce fondement correspond à la volonté de maintenir au cœur de notre modèle l'aspect continu du sens, c'est-à-dire le fait qu'une unité lexicale ne soit pas représentée par une liste de sens mais par un espace sémantique, et que retrouver le sens de cette unité ne corresponde pas à choisir un sens parmi sa liste de sens, mais à activer une zone de sens dans son espace sémantique. Nous pensons que la plupart des acteurs dans ce domaine seraient en accord avec cette idée théorique de l'aspect continu du sens. Le problème est que la plupart se heurtent à la difficulté de construire cet espace sémantique continu et sont donc obligés de se contenter de dictionnaires classiques ou adaptés au traitement automatique (WordNet), mais toujours construits sous forme de listes de définitions.

4.1. Continuité de sens

La linguistique pourrait être présentée comme un travail consistant à mettre en évidence les caractéristiques discrètes du langage, comme par exemple tenter de classer des éléments de la langue dans des catégories construites. Or cette volonté semble aller à l'encontre d'une des caractéristiques propres à toutes les langues : la gradualité. Par exemple, lorsque l'on cherche à définir des catégories sémantiques, comme des classes lexicales ('animé', 'humain', 'objet', etc.) ou des types de procès ('activité', 'accomplissement', 'achèvement', etc.), on se trouve toujours aux prises avec le même problème : s'il est très facile d'exhiber des exemples typiques présentant tous les traits qui caractérisent telle ou telle classe, il est aussi facile de trouver des exemples à la frontière de deux classes, pour lesquels il est nécessaire de distinguer différentes caractérisations, et d'introduire de nouveaux critères pour rendre compte de ces différences. Au fur et à mesure que le nombre de critères croît, les relations qu'ils entretiennent se complexifient, et leur combinatoire devient inextricable.

Il devient alors plus simple de raisonner dans un cadre continu. Ce dernier n'enlève en rien le fait qu'il existe différentes classes avec des exemples typiques. Et cela permet de placer les exemples qui sont à la frontière entre deux classes sur une échelle graduelle reliant les deux classes. Le cadre continu a aussi l'avantage de pouvoir être discrétisé alors que l'inverse n'est pas vrai. Autrement dit, le cadre continu permet aussi de traiter des tâches où la représentation discrète du sens serait mieux adaptée, par exemple pour une désambiguïsation grossière d'homonymes. Nous proposons d'illustrer quelques phénomènes de continuité à partir des trois verbes vedettes qui nous serviront d'exemples tout au long de cette thèse : *monter*, *compter*, et *jouer*.

4.1.1 Monter

Ce verbe a fait l'objet de plusieurs études en sémantique, notamment dans le cadre de travaux sur les verbes de mouvement ou de déplacement (Muller et Sarda, 1999 ; Sarda, 1999). Le TLFi en propose une définition contenant 73 nuances de sens, c'est-à-dire 73 sous-sens au plus bas de la structure hiérarchique de sa définition : c'est donc un verbe particulièrement polysémique[29]. A cela on doit ajouter quelques expressions telles que *la moutarde me monte au nez, monter au septième ciel, le vin monte à la tête*, soit en tout 10 pages de définitions. Voici les principales distinctions de sens proposées dans le TLFi :

[29] Pour retrouver les définitions du TLFi citées dans cette thèse : http://atilf.atilf.fr/tlf.htm .

MONTER, verbe

I. — *Emploi intrans.*

A. — [Le suj. désigne un être vivant]

1. Se déplacer dans un mouvement ascendant; s'élever dans un espace sans limites précises.

2. Se rendre dans un endroit situé plus haut que là où l'on se trouve.

3. S'élever pour se trouver plus haut, pour se rendre plus grand.

4. [En parlant de végétaux; souvent au passé] Parvenir au stade de la montaison.

5. Se déplacer pour se trouver dans un endroit stratégique ou à une place décisive pour l'action engagée.

6. Prendre place sur (un animal) ou dans (un véhicule).

7. [En parlant de provinciaux se rendant notamment dans la capitale] Aller du sud au nord.

8. [En parlant d'un ensemble de pers., d'une collectivité] **Monter à** ou, en emploi pronom., **se monter à**. S'élever au nombre de.

9. Accéder à un degré supérieur d'une hiérarchie.

10. Surenchérir. *JEUX DE CARTES.*

B. — [Le suj. désigne une chose]

1. Se déplacer dans un mouvement ascendant; s'élever dans l'air ou dans l'espace.

2. Partir d'un point bas pour aboutir à un point haut.

3. Gagner en hauteur (en parlant de constructions); croître en hauteur (en parlant de plantes).

4. [En parlant du mouvement ascendant des fluides] Atteindre une hauteur plus élevée, un niveau plus haut.

5. *MUS.* Passer du grave à l'aigu.

6. Augmenter, croître en quantité, en intensité, en valeur; atteindre telle ou telle valeur.

II. — *Emploi trans.*

A. — [*Monter* implique une notion d'élévation]

1. Parcourir en s'élevant, grimper, faire l'escalade de.

2. Prendre place sur le dos d'un animal; utiliser comme monture.

3. [En parlant du cheval et d'autres quadrupèdes mâles] Couvrir (la femelle).

4. Porter, mettre dans un endroit plus élevé, à plus grande hauteur.

B. — [*Monter* implique une notion d'agencement, d'assemblage, de mise au point]

1. Assembler les divers éléments d'un objet en vue de son utilisation.

2. *Au fig.* Organiser, mettre en oeuvre, combiner.

C. — [*Monter* implique une notion d'équipement] Pourvoir de tout le nécessaire.

Figure 4.1 : Structure de la définition de* monter *dans le TLFi

Notons que cette structure de définition est fondée sur des distinctions actantielles, avec une opposition entre emplois transitifs et intransitifs, et avec, au sein de la construction intransitive, une distinction entre un sujet « être vivant » et un sujet

« chose ». Il semble néanmoins que l'on puisse relier ces nuances de sens les unes aux autres par une « ressemblance de famille » à la Wittgenstein. Les sens (I.A.1.), (I.B.1.) et (II.A.1.) se rejoignent par la présence d'un déplacement ascendant d'un objet ou d'une personne. De même les sens (II.B.1.) et (II.B.2.) sont liés : lorsque l'on monte un projet, on retrouve la notion d'organisation mais aussi d'assemblage des divers composants du projet. Les sens (I.A.2.) et (I.B.5.) se distinguent par le domaine d'application, mais la notion de changement d'un niveau de référence à un niveau plus élevé est présente dans les deux cas.

On pourrait discuter de la manière d'organiser ces 73 nuances de sens par le TLFi. Notamment, la séparation entre emplois transitifs et intransitifs paraît dans certains cas très subtile. Cela implique de distinguer *monter la montagne* de *monter sur la montagne*. Certes, la notion d'accomplissement est présente dans le deuxième énoncé mais pas dans le premier, du moins pas nécessairement, mais cette distinction semble insignifiante par rapport à d'autres grands sens de *monter* tels que *surenchérir* et *prendre place sur un animal ou dans un véhicule*, qui sont pourtant tous deux dans la section (I.A.). On peut aussi retrouver des sens très différents avec une même construction verbale : *monter l'escalier / monter un projet / monter les valises.*

Nous pourrions alors penser que le sens de ce verbe est insensible à la construction verbale dans laquelle il se trouve et que les distinctions de sens du TLFi sont inappropriées. Nous verrons dans cette thèse qu'il est possible de rendre compte de la relation entre sens et construction même dans des cas aussi complexes que le verbe *monter*, et qu'il est possible d'exploiter ces relations pour la désambiguïsation. En ce qui concerne la structure hiérarchique proposée par le TLFi, elle peut être discutée, mais cela ne change pas le fait que, quelle que soit la structure hiérarchique choisie, il y aura toujours des ponts sémantiques entre les nuances de sens.

4.1.2 Compter

Ce verbe est moins polysémique, le TLFi propose 46 nuances de sens (le Grand Robert 2001 en propose 19), mais il est intéressant du point de vue des constructions verbales qu'il admet. Voici ci-dessous la structure hiérarchique de sa définition.

COMPTER, verbe.
I. — *Emplois trans.*
A. —
1. Déterminer une valeur ou une grandeur numérique par un calcul ou une suite de calculs, ou, le plus souvent, par une énumération, un dénombrement.
3. [Le compl. d'obj. fait partie d'un ensemble dénombrable] Inclure dans un ensemble, un total. Prendre en considération, tenir compte.
B. — *Au fig.* Déterminer une valeur ou une grandeur numérique future, la probabilité de réalisation d'une chose espérée.
1. [Constr. nom.] Prévoir, s'attendre à.
2. [Constr. verbale] Espérer, avoir l'intention de.
3. *Emploi abs.* **Compter sur qqn, qqc.**
II. — *Emplois intrans.*
A. — Être inclus dans un ensemble, un total.
B. — Avoir une certaine importance.

Figure 4.2 : Structure de la définition de* compter *dans le TLFi

Une nouvelle fois, le TLFi propose une opposition entre emplois transitifs et intransitifs. Ensuite, les deux sous-parties sont bien distinguées : I.A. implique la notion de dénombrement et I.B. la notion d'espérance. Malgré cela, il est encore une fois possible d'établir une continuité entre ces sens. L'énoncé *Je suis parti (...) pour le col de Bormes. Il faut compter six bonnes heures de chemin* (H. BOSCO, *Le Mas Théotime*, 1945, p. 329) qui se trouve dans la section (I.B.) illustre le chevauchement entre les sections (I.A.) et (I.B.). Dans cet énoncé, le verbe *compter* implique d'une part le calcul du temps de parcours et d'autre part un degré d'incertitude dans ce temps de calcul en précisant que ce n'est qu'un temps à espérer. On trouve aussi pour I.A.2. et II.A. la même notion « inclure dans un ensemble ».

Par rapport au verbe *monter*, *compter* est beaucoup plus dépendant de la construction verbale. *Compter pour quelque chose / compter sur quelque chose / compter quelque chose / compter avec quelque chose* sont des constructions qui

correspondent à des sens bien distincts, ce qui ne ressort pas clairement de la définition du TLFi. Par exemple, *compter sur quelque chose* est classé parmi les emplois transitifs en tant qu'emploi absolu[30]. Une fois de plus, on peut discuter ce choix, et il semble possible de considérer cette construction comme un emploi intransitif classique. C'est le cas dans la définition du Larousse 2003. Dans le Grand Robert 2001, cette construction est décrite dans une section regroupant des emplois intransitifs et des emplois transitifs indirects sans que la distinction soit vraiment précisée. Il n'est pas question ici de faire un choix entre ces trois structures ou encore d'en proposer une quatrième, nous voulons simplement montrer que la notion de continuité entre nuances de sens se retrouve aussi dans la description des différentes constructions du verbe.

4.1.3 Jouer

Le sens de *monter* est particulièrement sensible au lexique présent dans les arguments du verbe, le sens de *compter* est particulièrement sensible à sa structure argumentale. On peut replacer ces deux verbes aux deux extrémités d'une échelle graduelle sur laquelle se trouve toute une série de verbes, comme le verbe *jouer*, dont le sens dépend de manière plus ou moins forte d'une interaction entre lexique et syntaxe. Ce verbe, particulièrement polysémique, a fait l'objet de plusieurs études en sémantique (Nyckees, 1998 ; Cadiot, 1999 ; Leland, 2001 ; Romero-Lopes, 2002). Le TLFi décrit 116 nuances de sens auxquelles on doit ajouter 31 expressions telles que *jouer carte sur table, jouer de la prunelle, jouer franc jeu*, ce qui correspond en tout à 16 pages de définitions. Voici les principales distinctions de sens proposées :

[30] Un verbe transitif est dit en emploi absolu lorsqu'il s'emploie sans complément d'objet tout en restant implicitement transitif. Par exemple, dans la phrase *Elle a écrit hier*, le verbe *écrire* est en emploi absolu, alors que dans la phrase *Elle a écrit son article hier*, le verbe *écrire* est employé avec le complément d'objet *son article*.

JOUER, verbe

I. — [Le suj. désigne une pers.] **Qqn joue**

A. — Faire quelque chose pour se distraire, s'amuser.

B. — Faire quelque chose par jeu, par plaisanterie. *J'ai dit cela pour jouer.*

C. — Se livrer, avec une ou plusieurs autres personnes, à un jeu où l'on peut perdre ou gagner.

D. — Participer à un jeu, à une activité où l'argent et le hasard sont essentiels.

E. — Pratiquer un sport d'équipe, de façon professionnelle ou en amateur.

F. — *Emploi trans. indir.* Faire usage de quelque chose, avec plus ou moins d'adresse, de facilité.

G. — *Spécialement*

I. *MUS., emploi trans.* Exécuter, interpréter un morceau de musique, une œuvre musicale.

2. *THÉÂTRE, emploi trans.* [Le suj. désigne un acteur ou un ensemble d'acteurs : troupe, théâtre]

II. — [Le suj. désigne une chose] **Qqc. Joue**

A. — *Emploi intrans., littér.*

1. Se mouvoir avec aisance, comme au gré d'un jeu.

2. Produire de légers mouvements, qui entraînent des effets changeants

B. — *TECHNOL., emploi intrans.*

1. Se mouvoir, fonctionner dans un espace déterminé.

2. Fonctionner.

3. Se resserrer ou se dilater sous l'effet de causes naturelles.

Figure 4.3 : Structure de la définition de* jouer *dans le TLFi

Ici encore, les liens entre nuances de sens ne manquent pas. Par exemple, les sens (I.A.) et (I.B.) se rejoignent par la présence de la notion d'amusement, mais ils se distinguent par le fait que I.A. implique simplement une distraction (jouer à la marelle, avec une balle) alors que I.B. implique un objectif supplémentaire, cet objectif variant en fonction de l'objet d'amusement *(je jouais de lui avec une aisance merveilleuse, jouer avec sa réputation*). De même certains jeux (*jouer sur un cheval, jouer au poker*) permettent d'unifier les sens (I.C.) et (I.D.) puisque l'on retrouve la notion de gagner ou perdre, et la notion de hasard. On peut noter aussi que de tels jeux, qu'ils soient de hasard ou d'argent, que l'on puisse perdre ou gagner, n'excluent pas nécessairement la notion d'amusement présente dans les sens (I.A.) et (I.B.). La continuité est aussi présente entre les définitions du I (*quelqu'un qui joue*) et celles du II (*quelque chose qui joue*). Par exemple, (I.F.) et (II.A.) se distinguent par le domaine d'application, mais la

notion de maîtrise d'un élément dans un espace restreint donné se retrouve dans les énoncés le *poète joue avec les mots* et *la lumière joue avec le rideau*.

4.2. Fondement théorique

La notion d'espace continu est au cœur de notre modèle : nous avons voulu construire un modèle de désambiguïsation qui puisse traiter des représentations continues du sens et non des listes plus ou moins organisées de sens. Revenons d'abord sur le fondement théorique de cette approche : le modèle de la construction dynamique du sens (Victorri et Fuchs, 1996). L'idée de ce modèle est de pouvoir décrire et décomposer le processus de construction du sens, c'est-à-dire d'être capable, à partir d'un énoncé composé d'unités linguistiques, de modéliser la construction du sens de l'énoncé ainsi que la construction du sens des unités qui le composent, toute la difficulté étant de ne pas tomber dans un processus circulaire. En effet, le sens de l'énoncé est fonction du sens des unités qui le composent, et inversement le sens de ces unités dans cet énoncé est fonction du sens global de l'énoncé lui-même. On a donc affaire à un système qui obéit aux principes de base de la *Gestalttheorie* : le tout est plus que la somme de ses parties et « une partie dans un tout est autre chose que cette partie isolée ou dans un autre tout » (Guillaume 1979 : 23). Victorri et Fuchs (1996 : 41) proposent une mise en parallèle des systèmes complexes décrits par les Gestaltistes et de la construction du sens qu'il nous semble approprié de retranscrire intégralement :

> *« Comme l'ont bien montré les Gestaltistes, les systèmes sont régis par des règles d'optimalité : les interactions entre parties et tout conduisent le système dans un état, appelé une bonne forme, dans lequel un certain nombre de critères, qui dépendent de la nature des interactions, sont maximisés. Ainsi, en ce qui concerne la perception visuelle, domaine de prédilection des gestaltistes, ces critères ont été mis en évidence sous forme de lois : loi de simplicité, de proximité, de similarité, de prolongement, etc., et la forme perçue est celle qui satisfait le mieux possible ces lois. La bonne forme est donc l'aboutissement d'un processus dans lequel les éléments*

soumis à la perception jouent le rôle de contraintes, et ce processus vise à construire en fonction de ces contraintes une forme optimale. Pour en revenir au domaine de la langue, on peut considérer de la même manière que les unités qui composent un énoncé jouent aussi un rôle de contraintes, et que c'est l'aboutissement du processus de satisfaction de ces contraintes qui conduit au sens global de l'énoncé, conférant par là même un sens à toutes les unités qu'il contient. »

Autrement dit, la construction du sens de l'énoncé et des unités qui le composent est l'aboutissement d'interactions entre les contraintes imposées par les unités de l'énoncé. Toute la subtilité de cette conception réside dans l'aspect dynamique de ces interactions : le sens de chaque unité va être construit de manière progressive et simultanée jusqu'à atteindre un état stable. On voit émerger ici les principes d'un système dynamique qui vont nous permettre de rendre ce modèle de construction du sens non circulaire.

Dans ce processus, les contraintes de chaque unité correspondent à l'ensemble des sens que cette unité peut prendre. Nous appelons cet ensemble *le potentiel de sens de l'unité*. Afin de respecter l'aspect continu du sens, l'idée est ici de représenter ce potentiel de sens par un espace géométrique, c'est-à-dire un espace dans lequel un sens est représenté par une région. Nous assimilons un sens à une région plutôt qu'à un point afin de pouvoir rendre compte du fait que les sens d'une unité dans deux énoncés différents peuvent se recouper, c'est-à-dire partager une partie commune sans pour autant être parfaitement identiques. D'une manière plus générale, la taille et la forme de ces régions servent à modéliser des caractéristiques importantes des sens : une région étendue correspond à un sens plus indéterminé, une région étroite à un sens très précis, une région non connexe (constituée de plusieurs parties disjointes) à une ambiguïté, etc. Calculer le sens de l'unité dans un tel espace consistera alors à activer une région de cet espace. Reste à savoir comment construire un tel espace. C'est ce que nous présentons maintenant.

4.3. Développement logiciel

4.3.1 Construction d'un espace sémantique

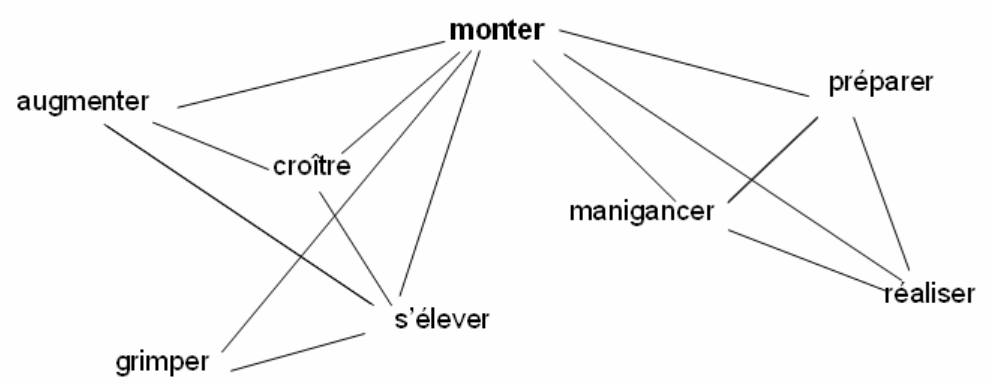
Certains tenants de la LSA (*Latent Semantic Analysis*) proposent la construction de « *Latent Semantic Spaces* » (Hofmann, 1999). La construction de ces espaces sémantiques est fondée sur des relations de cooccurrence, le principe étant que deux mots ayant des cooccurrents proches sont proches dans l'espace sémantique. Nous sommes en accord avec ce principe ; néanmoins ces espaces sont dépendants d'un domaine particulier de la langue, voire d'un corpus. Notre objectif, du moins théorique, est de construire un espace permettant de rendre compte, en langue générale, du plus grand nombre de sens possible d'une unité lexicale.

Pour cela nous utilisons le dictionnaire électronique des synonymes (D.E.S.) du français du laboratoire CRISCO (www.crisco.unicaen.fr). Outre le fait d'être numérisé, ce dictionnaire a l'avantage d'être construit à partir de sept dictionnaires classiques (Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse et Robert) dont ont été extraites les relations synonymiques. Il contient 49 000 entrées et 396 000 relations synonymiques. Etant donné la diversité des dictionnaires exploités, nous pouvons estimer que ce dictionnaire permet une représentation assez complète des relations synonymiques de la langue française[31]. L'idée est alors de définir chaque sens en terme de synonymes remplaçables, tout en gardant à l'esprit qu'un seul synonyme ne suffit pas à définir un sens puisque les synonymes d'un verbe polysémique peuvent aussi être polysémiques.

Le D.E.S. ne se contente pas de donner la liste des synonymes d'une unité, il en construit le graphe. Dans ce graphe, chaque sommet est une unité, et deux unités sont mises en relation si elles sont synonymes. Ce graphe est très complexe et semble

[31] Nous sommes conscient que cette représentation des relations de synonymie est une représentation statique. Nous ne traiterons pas ici de l'évolution des relations de synonymie dans le temps.

confirmer l'idée qu'un simple synonyme n'est généralement pas suffisant pour définir un sens précis d'une unité. Considérons par exemple le verbe *monter*. Le DES en propose 107 synonymes[32], et voici un sous-ensemble du graphe contenant ce verbe :



Le verbe *s'élever* illustre bien le fait qu'un seul synonyme n'est pas suffisant pour définir un sens précis. Il est à la fois synonyme de *grimper* et de *augmenter*, ce qui correspond à deux sens distincts de *monter* : le sens lié à l'augmentation du niveau d'un élément (eau, prix, etc.), et celui lié à l'ascension d'une montagne, d'un sommet. Or les points de notre espace sémantique doivent correspondre à des sens précis de l'unité.

L'idée est alors d'introduire la notion de clique (Ploux et Victorri, 1998). Une clique est un sous-graphe complet maximal, c'est-à-dire un ensemble le plus grand possible de sommets tous reliés deux à deux. C'est donc ici un ensemble d'unités

32 Liste des synonymes de *monter : affluer ; agencer ; ajuster ; aller ; appareiller ; arriver ; ascensionner ; assembler ; atteindre ; attirer ; augmenter ; aviver ; bâtir ; chevaucher ; chiffrer ; combiner ; constituer ; coudre ; couvrir ; croître ; créer ; disposer ; doubler ; dresser ; échafauder ; élever; embarquer ; enchatonner ; enchâsser ; enfourcher ; enlever ; entrer ; escalader ; établir ; exciter ; exhausser ; fabriquer ; faire ; forcer ; franchir ; grandir ; gravir ; grimper ; grossir ; hausser ; hisser ; installer ; jouer ; lever ; machiner ; majorer ; manigancer ; mettre en scène ; mijoter ; mitonner ; monter ; nouer ; organiser ; ourdir ; parvenir ; percer ; planter ; porter ; pourvoir ; prendre ; procurer ; progresser ; préparer ; redoubler ; rehausser ; relever ; remonter ; renchérir ; revaloriser ; réaliser ; réussir ; s'accentuer ; s'accoupler ; s'accroître ; s'amplifier ; s'embarquer ; s'engouffrer ; s'envoler ; s'installer ; s'intensifier ; s'échapper ; s'édifier ; s'élever ; saillir ; se bâtir ; se construire ; se débourgeoiser ; se débourrer ; se guinder ; se hausser ; se hisser ; se monter ; se percher ; sertir ; servir ; soulever ; surhausser ; surélever ; tisser ; tramer ; tresser ; voler.*

synonymes deux à deux. La portion du graphe de *monter* que nous avons représentée ici contient 3 cliques, correspondant chacune à un sens précis de *monter*[33] :

- *augmenter, croître, s'élever, monter*
- *grimper, s'élever, monter*
- *préparer, réaliser, manigancer, monter*

Nous faisons l'hypothèse que chaque clique représente un sens très précis de l'unité que l'on peut associer à une région suffisamment petite de l'espace sémantique pour que l'on puisse l'assimiler à un point. Notre espace sémantique est une projection en deux dimensions du nuage formé par les cliques dans l'espace multidimensionnel engendré par les synonymes de l'unité lexicale considérée. Il est muni de la métrique du χ^2. Plus précisément, appelons u_1, u_2, …, u_n les synonymes de l'unité, c_1, c_2, …, c_p les cliques associées et posons $x_{ki} = 1$ si $u_i \in c_k$ et $x_{ki} = 0$ si $u_i \notin c_k$. La distance $d(c_k, c_l)$ entre deux cliques est alors définie de la façon suivante (cf. Ploux et Victorri, 1998. Victorri, 2002) :

$$d^2(c_k, c_l) = \sum_{i=1}^{n} \frac{x}{x_{\bullet i}} \left(\frac{x_{ki}}{x_{k\bullet}} - \frac{x_{li}}{x_{l\bullet}} \right)^2$$

$$\text{avec } x_{\bullet i} = \sum_{j=1}^{p} x_{ji} \,, \quad x_{k\bullet} = \sum_{i=1}^{n} x_{ki} \,, \text{ et } \quad x = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ji} \,.$$

Autrement dit, cette distance possède les deux caractéristiques suivantes. D'une part, chaque synonyme, en tant que vecteur de base de l'espace, intervient dans le calcul avec un « poids » plus faible si le synonyme est présent dans un grand nombre de cliques : les synonymes qui sont les moins spécifiques jouent un rôle moins important dans la discrimination des sens de l'unité. D'autre part, les coordonnées de chaque clique sont divisées par le nombre d'éléments de la clique : le point représentant la

[33] Le verbe *monter* possède au total 143 cliques (cf. annexe *G.1. pour avoir la liste intégrale).*

clique n'est donc pas un sommet de l'hypercube mais il est d'autant plus proche de l'origine que la clique correspondante comporte plus de synonymes.

Notre hypercube possède autant de dimensions qu'il y a de synonymes, 107 pour le verbe *monter*. En sélectionnant les deux dimensions qui dispersent le plus l'espace des cliques, à l'aide d'une AFC (analyse factorielle des correspondances), nous obtenons alors une représentation graphique en 2D de toutes les cliques de synonymes (cf. Figure 4.4, cf. annexe A.1. pour une visualisation 3D de l'espace sémantique de *monter* avec les trois premières composantes de l'AFC). Cette représentation aura rempli son rôle si elle met en évidence les sens importants du verbe et surtout si elle rend compte des différentes nuances de sens et des ponts sémantiques entre sens.

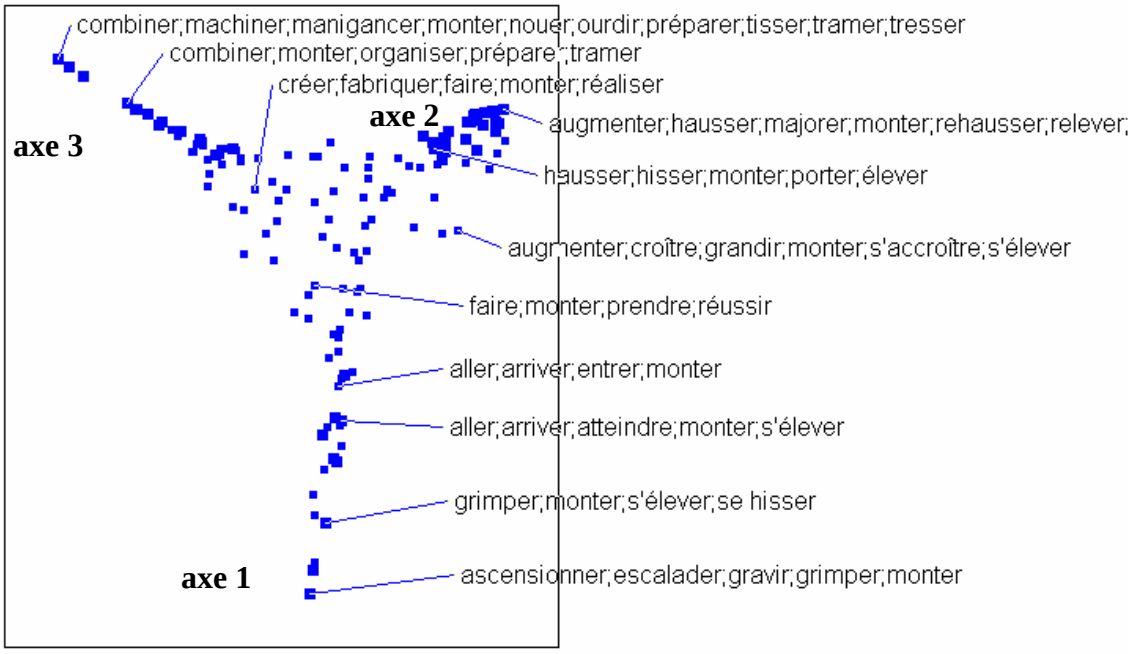


***Figure 4.4 : Espace sémantique du verbe* monter**

Rappelons que dans l'espace représenté figure 4.4, chaque point est une clique et la distance entre cliques est fondée sur la métrique du χ^2, ce qui revient à fonder cette distance sur deux principes :

- Plus deux cliques ont de synonymes en commun, plus elles sont proches
- Plus un synonyme est omniprésent dans les cliques (comme par exemple le mot vedette qui appartient à toutes les cliques), moins il est influent pour la distance.

L'espace sémantique obtenu semble répondre à nos attentes. Les trois axes dominants correspondent à trois sens important de *monter* : l'axe 1 représente le sens « *gravir, escalader une montagne* », l'axe 2 représente le sens « *augmenter, relever les prix* » et l'axe 3 représente le sens « *organiser, préparer un projet* ». On pourrait même nommer ces axes à l'aide des sections de sens proposées par le TLFi (cf. § 4.1.1) :

- Axe 1 : être vivant se déplaçant dans un mouvement ascendant/se rendant dans un endroit plus haut que là où il se trouve (sections I.A.1. et I.A.2.)
- Axe 2 : chose qui augmente en quantité, en valeur (section I.B.6.)
- Axe 3 : assembler les divers éléments d'un objet/organiser, mettre en œuvre (sections II.B.1. et II.B.2.).

L'axe 1 illustre très bien la manière de représenter les nuances de sens. Reprenons les quatre cliques suivantes :

ascensionner, escalader, gravir, grimper, monter

grimper, monter, s'élever, se hisser

aller, arriver, atteindre, monter, s'élever

aller, arriver, entrer, monter

Elles véhiculent toute la notion « être vivant se déplaçant dans un mouvement ascendant/se rendant dans un endroit plus haut que là où il se trouve » de l'axe 1. On peut noter cependant un changement progressif du sens allant d'un déplacement où la

direction verticale est mise en relief avec *ascensionner, escalader, gravir, grimper, monter*, à un déplacement où elle ne l'est plus avec *aller, arriver, entrer, monter*, les deux autres cliques correspondant à des cas intermédiaires. Et de nuances de sens en nuances de sens, nous sommes passé de *monter la montagne* à *monter à Paris* qui sont deux sens bien distincts dans le TLFi : respectivement les sections (II.A.1.) et (I.A.7.). Nous pensons pouvoir dire que cet espace sémantique répond à nos attentes concernant la mise en évidence tant des sens importants du verbe *monter* que des ponts sémantiques entre sens distincts.

4.3.2 Espace sémantique du verbe compter

Etudions maintenant le verbe *compter*. Rappelons que le TLFi en propose 46 nuances de sens. Le D.E.S. énumère 62 synonymes[34] et 67 cliques (cf. annexe G.2. pour avoir la liste complète des cliques contenant *compter*), c'est-à-dire 67 nuances de sens (cf. § 4.1.2). Calculons maintenant l'espace sémantique de *compter* (avec la même méthode que pour le verbe *monter*).

[34] Liste des synonymes de *compter : penser ; estimer ; considérer ; peser ; espérer ; regarder ; évaluer ; apprécier ; croire ; examiner ; présumer ; supputer ; envisager ; prendre ; calculer ; escompter ; mesurer ; réputer ; comprendre ; contenir ; dénombrer ; attendre ; s'attendre ; se flatter ; songer ; supposer ; tenir pour ; importer ; inclure ; inventorier ; nombrer ; projeter ; se promettre ; se proposer ; tabler ; énumérer ; chiffrer ; englober ; entendre ; introduire ; marquer ; posséder ; présenter ; recenser ; se targuer ; entrer en ligne de compte ; avoir l'intention ; dater ; décompter ; fonder ; s'appuyer ; être important ; avoir pour certain ; avoir pour sûr ; computer ; exister ; facturer ; payer ; précompter ; se ranger ; épargner.*

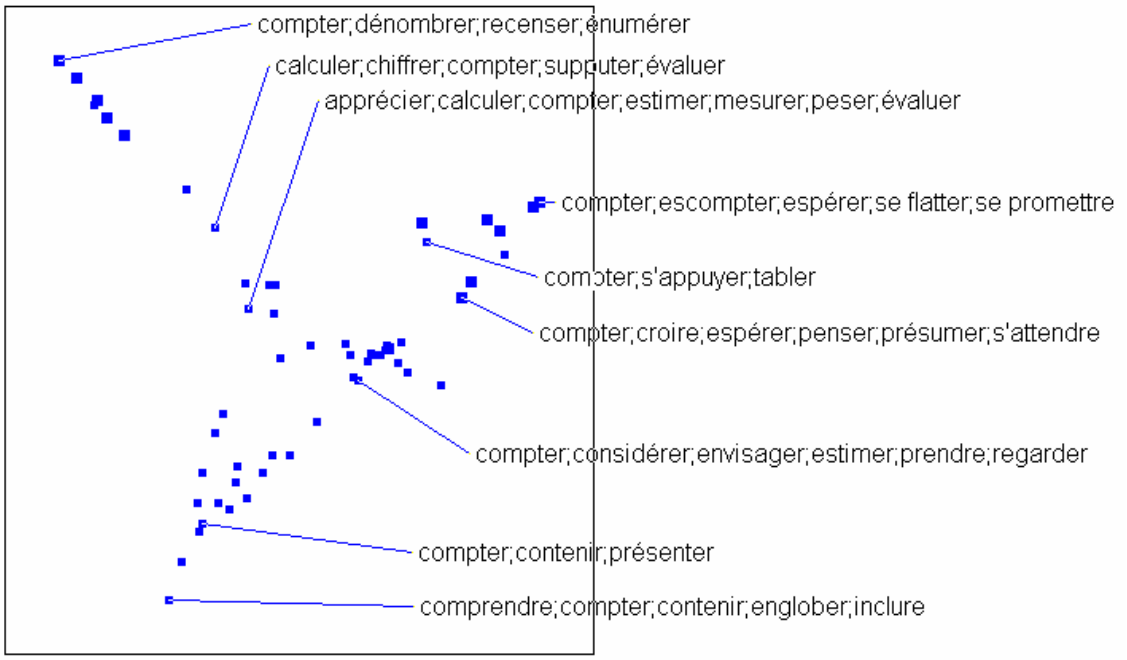


***Figure 4.5 : Espace sémantique du verbe* compter**

On peut noter, à l'aide de la figure 4.5 (cf. annexe A.2. pour une visualisation 3D de l'espace sémantique de *compter*) que des « branches de sens » se démarquent très nettement. Notre modèle met en évidence trois principales « branches » qui sont :

- en bas les sens proches de *comprendre*, *compter*, *contenir*, *englober*, *inclure* ;

- en haut à droite : *attendre*, *compter*, *escompter*, *espérer*, *se promettre* ;

- en haut à gauche : *compter*, *dénombrer*, *recenser*, *énumérer*.

Outre le fait de faire émerger les sens forts de cette unité, l'intérêt d'une telle représentation est de garder intacte l'aspect continu du sens. On peut en effet aller progressivement d'un sens fort à un autre. Par exemple, on peut aller de {*compter dénombrer, recenser, énumérer*} à {*compter considérer, envisager, penser, songer*} en passant progressivement par {*calculer, chiffrer, compter, supputer, évaluer*} puis par {*apprécier, calculer, compter, estimer, mesurer, peser, évaluer*}.

4.3.3 Espace sémantique du verbe **jouer**

Prenons le verbe *jouer*. Le D.E.S. en propose 91 synonymes[35] et 98 cliques (cf. annexe G.3. pour avoir la liste complète des cliques contenant *jouer*). Etudions l'espace sémantique que nous avons calculé (cf. figure 4.6 ; cf. annexe A.3. pour une visualisation 3D de l'espace sémantique de *jouer*). On peut noter trois branches principales de sens : en haut les sens tournant autour de « *jouer, miser, parier, ponter* », en bas à gauche « *badiner, folâtrer, jouer, plaisanter, s'amuser* », puis en bas à droite « *contrefaire, feindre, imiter, jouer, simuler* ». Encore une fois, l'aspect continu du sens est respecté puisque l'on peut glisser progressivement de *copier, imite, mimer*, à *pratiquer, faire, exécuter* en passant par *reproduire, simuler* puis *incarner, représenter, interpréter.*

[35]Liste des synonymes de *jouer : abuser, affecter, agir, agiter, amuser, attaquer, aventurer, avoir du jeu, badiner, batifoler, berner, blaguer, boursicoter, compromettre, contrefaire, copier, coulisser, crier, créer, donner, duper, enlever, exposer, exécuter, faire, feindre, feinter, figurer, flamber, flouer, folâtrer, fonctionner, gauchir, gondoler, gratter, hasarder, imiter, incarner, influer, interpréter, intervenir, jongler, jouer, manier, marcher, massacrer, mentir, mettre en scène, mimer, miser, monter, mystifier, parier, passer, pianoter, pincer, plaisanter, plastronner, ponter, poser, pratiquer, racler, railler, refaire, remuer, reproduire, représenter, rire, risquer, rouler, s'amuser, s'entraîner, s'exercer, s'ébattre, s'ébrouer, se divertir, se faire entendre, se mouvoir, secouer, simuler, singer, sonner, souffler, spéculer, taquiner, tenir, toucher, tourner, tripoter, tromper, user.*

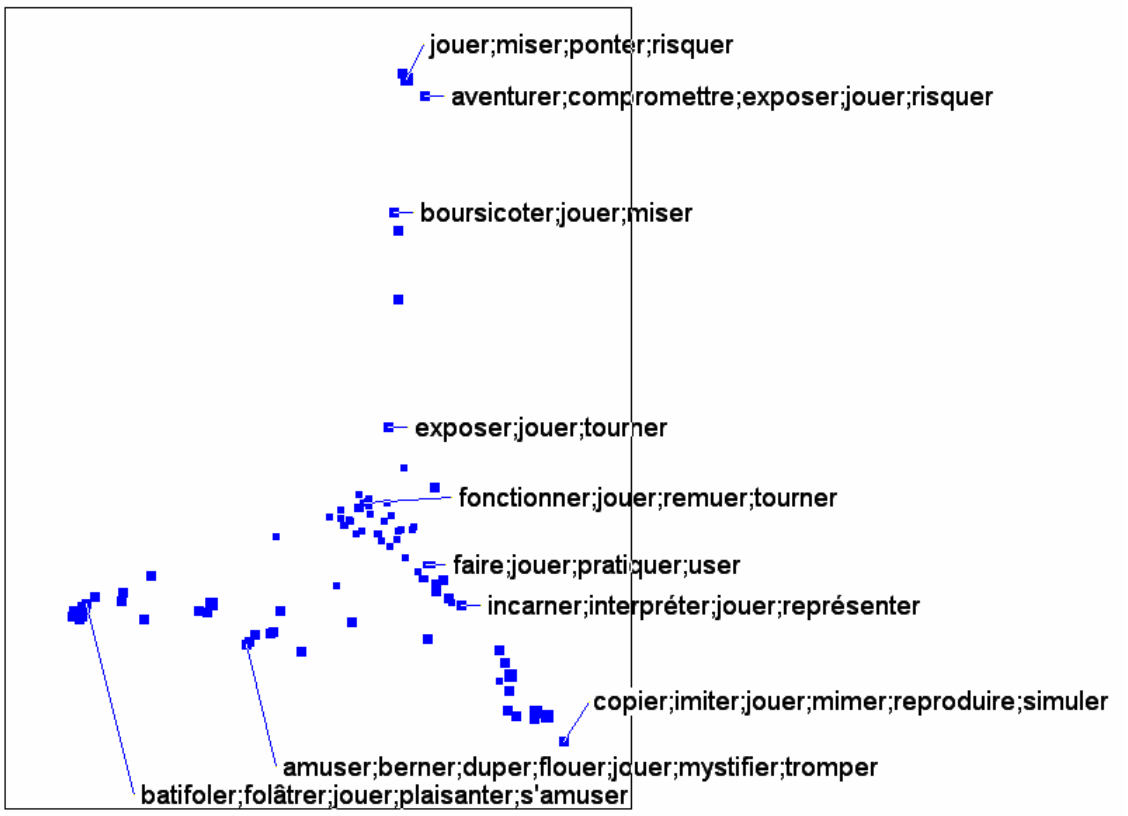


***Figure 4.6 : Espace sémantique du verbe* jouer**

4.3.4 Région sémantique associée à un synonyme

Nous venons de voir quelques nuances de sens et leurs comportements dans l'espace sémantique. Une nuance de sens est facilement identifiable dans l'espace puisqu'elle est représentée par un point. Un sens plus flou ou plus général doit correspondre à une région de l'espace. Nous avons proposé un modèle dans le paragraphe 4.2 basé sur le calcul de fonctions potentielles permettant d'activer une région de sens de cet espace. Avant d'appliquer cette méthode à la désambiguïsation en contexte, nous proposons de l'expérimenter sur les synonymes du mot vedette : il s'agit donc d'associer une fonction à chaque synonyme du mot vedette afin de déterminer si celle-ci permet de visualiser la région de l'espace sémantique dans laquelle la relation de synonymie entre le mot vedette et le synonyme considéré est pertinente. Cette fonction est calculée sur l'ensemble des cliques en donnant un poids égal à 1 aux cliques contenant ce synonyme, et un poids égal à –0.1 aux cliques ne le contenant pas.

Appelons $u_1,u_2,...,u_m$ les synonymes et $c_1,c_2,\ldots,c_c$ les cliques. La valeur de la fonction associée au synonyme u_j au point de coordonnées (x;y) est donnée par :

$$f_j(x,y) = \max\left(0, \sum_{k=1}^{c} a_{jk} \times e^{-\frac{(x_k-x)^2+(y_k-y)^2}{\delta^2}}\right)$$

où (x_k,y_k) sont les coordonnées du point représentant la clique c_i dans l'espace sémantique,

a_{jk}=1 si u_j appartient à la clique c_k, -0.1 sinon

et $\delta = \frac{\max(dx,dy)}{10}$ avec $dx = xsup - xinf$ et $dy = ysup - yinf$

$$xsup = \max_{k=1}^{c}(x_i) \text{ et } xinf = \min_{k=1}^{c}(x_i)$$

$$ysup = \max_{k=1}^{c}(y_i) \text{ et } yinf = \min_{k=1}^{c}(y_i)$$

A titre d'exemple, on trouvera en figure 4.7 les régions activées respectivement par *manigancer*, *augmenter* et *s'élever*. Ces trois synonymes illustrent bien les différents cas de figure pouvant se présenter. *Manigancer* correspond à un sens très précis de *monter* (préparer, faire des manigances. *Moi, qui ai demandé comment les élus de Paris qui étaient en même temps députés à Versailles allaient manigancer leur petite affaire* (VALLÈS, *J. Vingtras,* L'insurgé, 1885, p. 293). La région activée par la fonction de ce synonyme correspond effectivement à un sens très précis puisqu'elle ne comprend que trois cliques dont la clique {*combiner, machiner, manigancer, mijoter, monter, ourdir, préparer, tramer*}.

A l'inverse, la région activée par le synonyme *augmenter* est beaucoup plus étendue et semble correspondre à une indétermination de sens. En effet, *augmenter* en tant que synonyme de *monter* peut vouloir dire *rehausser, relever, revaloriser* comme dans *monter/augmenter les prix*, mais aussi *grandir, s'élever, croître* comme dans *le*

niveau de l'eau monte/augmente. Cette indétermination ne provient pas des expressions lexicales *prix* et *niveau de l'eau* puisqu'il est facile d'inverser les rôles : *monter/augmenter le niveau de l'eau* et *les prix montent/augmentent*. Autrement dit, il y a bien indétermination et elle ne provient pas du lexique mais du synonyme et de ses constructions, avec la possibilité que l'action d'augmenter soit plus ou moins maîtrisée par une entité extérieure sans que cela nécessite d'être précisé.

Enfin, le synonyme *s'élever* illustre un cas d'ambiguïté. *S'élever*, en tant que synonyme de *monter*, peut vouloir dire *s'accroître, augmenter, croître* comme dans *les prix s'élèvent/montent* ou bien *grimper, se hisser* comme dans *s'élever/monter en haut de la montagne*. Contrairement au cas du verbe *augmenter*, les deux sens possibles ne sont pas compatibles. Autrement dit, *monter* au sens de *s'élever* reste encore ambigu, et notre modèle permet de rendre compte de cette ambiguïté en faisant ressortir deux bassins distincts.

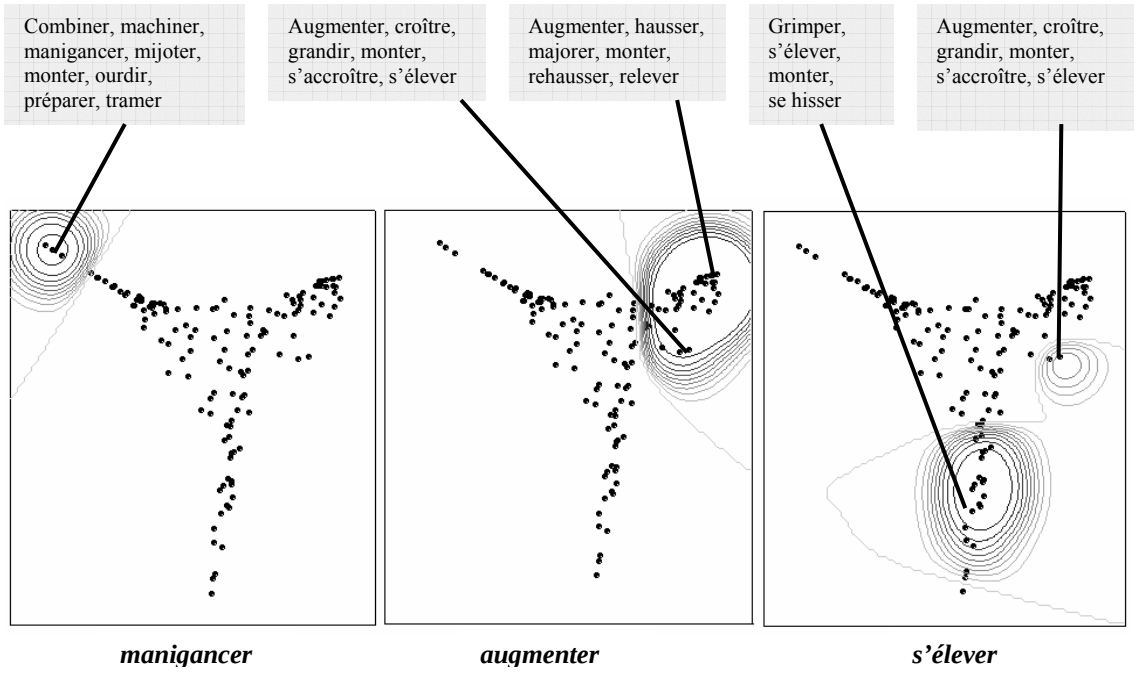


***Figure 4.7 : Fonctions potentielles des synonymes* manigancer, augmenter *et* s'élever *sur l'espace sémantique du verbe* monter**

5. Une méthode de calcul du sens

Sur la base du modèle de représentation du sens présenté au chapitre précédent, un modèle théorique de construction dynamique du sens avait été élaboré il y a une dizaine d'années (Victorri et Fuchs, 1996). Mais ce n'est que très récemment que ce modèle a été rendu opérationnel grâce à l'utilisation de gros corpus qui ont permis de calculer effectivement l'influence d'un élément du co-texte sur le sens d'une unité polysémique. Ce travail a été mené par une équipe du laboratoire LATTICE composée essentiellement de Fabienne Venant et moi-même sous la direction de Bernard Victorri.

Dans ce chapitre, nous commencerons par présenter le modèle théorique de construction du sens avant de décrire la méthode générale de calcul qui a été mise au point par notre équipe. Nous présenterons aussi brièvement les corpus que nous avons utilisés dans nos expérimentations sur les verbes. Dans les chapitre suivants, nous présenterons différentes études sur la polysémie verbale : il s'agit là d'un travail entièrement personnel qui s'inscrit dans le projet collectif de notre équipe. Ce travail a aussi joué un rôle moteur dans un projet soutenu par l'ILF (Institut de linguistique française : fédération de laboratoires du CNRS), intitulé « Polysémie verbale : le rôle de la construction syntaxique », dirigé par Jacques François et impliquant trois équipes : une équipe du CRISCO de Caen, dirigée par Jacques François et Jean-Luc Manguin, une de l'ERSS de Toulouse, dirigée par Didier Bourigault et notre équipe du LATTICE.

5.1. Le modèle théorique de construction dynamique du sens

Nous avons déjà indiqué que le cadre des systèmes dynamiques était particulièrement bien adapté pour modéliser le principe de compositionnalité Gestaltiste. Victorri et Fuchs (1996) ont proposé de représenter l'influence des éléments

du co-texte sur le sens d'une unité polysémique par la donnée d'une fonction potentielle sur l'espace sémantique associé à l'unité (cf. figure 5.1). Cette fonction potentielle définit une dynamique sur cet espace et les bassins d'attracteurs correspondent aux différents sens possibles de l'unité dans l'énoncé considéré. Plus précisément, les valeurs du potentiel inférieures à un certain seuil, appelé seuil d'admissibilité, déterminent une région de l'espace sémantique qui représente le sens de l'unité dans l'énoncé.

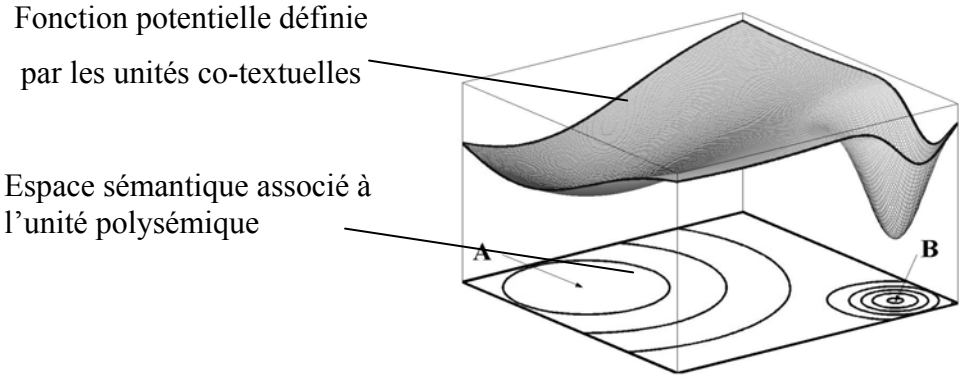
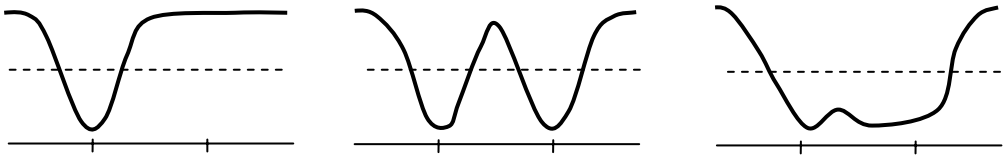


Figure 5.1 : Représentation d'une fonction potentielle sur un espace sémantique bidimensionnel

Le grand intérêt de cette modélisation, outre le fait qu'elle est bien adaptée à la conception continue du sens que nous avons défendue au chapitre précédent, c'est qu'elle permet de représenter différents cas de figure interprétatifs suivant la forme de la fonction potentielle, comme on peut le voir figure 5.2. Une fonction potentielle ne contenant qu'un bassin peu étendu correspondra à un sens précis (5.2a), deux bassins peu étendus correspondront à des cas d'ambiguïté (5.2b) et un bassin plus étendu correspondra aux cas d'indétermination (5.2c)[36]. Examinons plus en détail ce que l'on entend par ces différents cas de figure.

[36]On trouve des représentations analogues dans Sadock (1986).



(5.2a) sens précis (5.2b) ambiguïté-alternative (5.2c) indétermination

Figure 5.2 : Les différents cas de figure interprétatifs (espace unidimensionnel)

Le premier cas de figure correspond aux cas où l'énoncé détermine de façon très claire un sens précis, bien établi, prototypique de l'unité. Par exemple, pour le verbe *monter*, il n'y a aucune hésitation des locuteurs pour définir son sens dans des énoncés tels que :

(1) *Pierre est monté à l'échelle pour réparer la gouttière*.

(2) *Le groom va monter les valises dans votre chambre.*

(3) *La cote de cette action est montée brutalement dans la journée d'hier.*

(4) *Pierre a monté le meuble en kit qu'on avait acheté la semaine dernière*.

(5) *Pierre a monté une expédition en Amazonie.*

On a, de manière très claire, le sens d'élévation dans les exemples (1) et (2), celui d'augmentation dans le (3), d'assemblage dans le (4) et enfin d'organisation dans le (5). Dans le modèle, cela correspond à un bassin étroit et profond, ce qui signifie que la valeur précise du seuil d'interprétation, indiqué sur la figure par une ligne pointillée, a peu d'importance (si elle reste dans des limites raisonnables, bien sûr) : le sens sera toujours la même petite région. C'est ainsi que l'on traduit le fait que l'on aura une très grande homogénéité des jugements des locuteurs sur ces emplois.

Le deuxième cas de figure, celui de l'ambiguïté, correspond à la situation où l'énoncé dans son ensemble reste ambigu : il ne contient pas suffisamment

d'information pour que l'on puisse trancher entre deux sens (ou plus) nettement distincts. C'est le cas, par exemple, des énoncés suivants :

(6) *Pierre est en train de monter l'étagère dans la chambre*

(7) *Pierre a monté tous les dossiers de financement de ces dernières années*

Dans l'énoncé (6), on ne peut pas choisir entre les sens d'élévation[37] et d'assemblage. De même, l'énoncé (7) aura le sens d'élévation s'il s'agit des dossiers en tant qu'objets physiques, et celui d'organisation s'il s'agit du contenu des dossiers. Dans le modèle, on a deux bassins, séparés par un col très élevé. De même que dans le premier cas de figure, cela veut dire que la position précise du seuil d'admissibilité ne compte pas, mais cette fois le sens est représenté par l'union de deux régions non connexes. Pour les locuteurs, tout dépend du type d'interprétation qu'ils ont à fournir : la plupart du temps, ils seront cantonnés à une seule région, c'est-à-dire qu'ils ne « verront » qu'une interprétation, qui dépendra du contexte global et de leur état d'esprit. Mais dans certains cas, notamment si on leur demande explicitement de rechercher les différents sens de la phrase, ils pourront donner les deux interprétations en « explorant » entièrement l'espace sémantique. Il faut noter le caractère artificiel de ce cas de figure : c'est parce que l'on analyse une phrase isolée de son contexte que ces ambiguïtés se manifestent. Dans un texte, le co-texte dans son ensemble (c'est-à-dire en prenant en compte les phrases qui précèdent) suffit généralement à lever l'incertitude, sauf dans des cas très spécifiques, quand ce double sens est intentionnel, par exemple dans le cas d'un jeu de mots.

Le troisième cas de figure, celui de l'indétermination, est d'une certaine manière le plus intéressant. En effet, il regroupe tous les cas où deux sens habituels du mot (ou plus de deux) sont co-présents dans le sens véhiculé par l'énoncé. Les deux sens en

[37] Avec d'ailleurs deux sous-sens possibles dans ce cas : Pierre peut être dans les escaliers en train de transporter l'étagère, ou simplement en train de la rehausser de quelques centimètres de sa position initiale dans la chambre.

question ne s'opposent pas, on n'a pas à choisir entre eux. Au contraire, ils se complètent, c'est-à-dire qu'ils permettent de construire un sens plus riche parce qu'il « joue » sur une partie plus importante du sémantisme de l'unité. Pour prendre un exemple, prenons le passage suivant de *l'Etranger* de Camus (cité par le TLFi) :

Il [*un journaliste*] *m'a dit qu'il espérait que tout irait bien pour moi (...) et il a ajouté: « Vous savez, nous avons monté un peu votre affaire. L'été, c'est la saison creuse pour les journaux. Et il n'y avait que votre affaire et celle du parricide qui vaillent quelque chose »*.

Dans cet exemple, les sens d'augmentation, d'assemblage et d'organisation co-existent. Les journalistes ont donné à cette affaire plus d'importance qu'elle n'en avait, mais on peut aussi dire qu'ils ont en partie fabriqué l'affaire en question. Ce double sens de *monter* est beaucoup plus fréquent qu'on ne pourrait le penser. Notamment, l'expression *monter en épingle* joue parfaitement sur ce double sens : cette expression provient, d'après le TLFi, de la référence aux épingles de parure (épingles de cravates et de chapeaux) comme d'ailleurs *tiré à quatre épingles*. C'est donc au départ le sens d'assemblage qui est utilisé. Mais cela n'empêche pas le TLFi de classer ce sens dans la partie A, qui regroupe tous les emplois transitifs du verbe dans les sens d'élévation et d'augmentation (cf. chapitre précédent, §4.1.1).

Il ne faudrait pas croire que ces cas de figure d'indétermination ne touchent que des emplois abstraits, « figurés ». Pour donner un exemple dans un domaine très concret, la cuisine, *monter une mayonnaise* joue aussi sur deux sens : assemblage (on confectionne la mayonnaise à partir des ingrédients) et élévation (le produit occupe plus de place dans le bol : c'est d'ailleurs ainsi que le comprend le TLFi, puisqu'il classe *monter une mayonnaise, des blancs en neige, une sauce au beurre* dans la partie A).

On peut aussi considérer comme indétermination un cas de figure un peu différent mais difficile à distinguer du précédent dans la pratique : celui où on a affaire à un sens intermédiaire entre deux sens bien établis. Par exemple *monter un film* est intermédiaire entre les sens d'assemblage et d'organisation. D'une part, on met bout à

bout des morceaux de film qui existent déjà, c'est donc bien de l'assemblage, mais contrairement à ce qui se passe pour un meuble en kit, le « mode d'emploi » n'est pas complètement fixé, on sélectionne, on aménage et on réaménage au point de changer sensiblement le scénario initial : il y a donc bien un sens d'organisation. Il est difficile de dire la part d'assemblage et la part d'organisation parce que cela change d'une fois à l'autre. On peut faire le même genre de remarque pour *monter une pièce de théâtre* ou un spectacle en général.

Les cas d'indétermination ne sont donc pas des exceptions, un peu artificielles, comme les ambiguïtés alternatives. Au contraire, on peut même dire que ce sont eux qui justifient que la polysémie se maintienne dans la langue car c'est pour ces cas-là que la polysémie est vraiment utile et efficace dans la communication. Pour les locuteurs, il n'est pas difficile de comprendre ce que le mot veut dire dans un cas d'indétermination. En revanche, il est très difficile pour eux de dire exactement à quel sens général du mot cet emploi se rattache. Les jugements de différents locuteurs peuvent beaucoup diverger : certains rattachant l'emploi sans problème à un des sens, d'autres à l'autre, d'autres encore aux deux, et d'autres enfin à aucun des deux, estimant qu'il s'agit d'un nouveau sens, qu'il faut répertorier en tant que tel. Cela explique à notre avis pourquoi il est si difficile d'obtenir un accord interjuge satisfaisant pour évaluer les systèmes de désambiguïsation automatiques, comme nous l'avons signalé (cf. chapitre 3, § 3.2.3)[38]. Il faut noter que cela se produit même si on ne veut qu'une granularité grossière : pour *monter une mayonnaise*, c'est entre les deux sens les plus éloignés de *monter* qu'il y a indétermination…

Le modèle cherche à rendre cette incertitude de jugement dans les cas d'indétermination par un bassin très large et assez plat, couvrant les deux sens en question, le fond du bassin étant plus ou moins accidenté. Cela veut dire que si le seuil d'interprétation est assez élevé (comme sur la figure 5.2c), le sens est représenté par une

[38] Voir aussi Véronis (1998, 2004).

grande région connexe. En revanche, si le seuil s'abaisse suffisamment (c'est-à-dire si l'on est très exigeant), on peut passer à des petites régions non connexes, correspondant à des divergences de jugement entre locuteurs. Ainsi le modèle permet bien de rendre compte de la variabilité des interprétations dans ce cas de figure. Il faut d'ailleurs remarquer qu'on peut, en plus, avoir des cas de figure intermédiaires entre le cas de l'ambiguïté-alternative et celui de l'indétermination : s'il y a deux bassins, mais que la hauteur du col n'est pas très grande, une partie des locuteurs jugera que l'on a affaire à une ambiguïté (ceux qui auront un seuil d'admissibilité très bas) et les autres que l'on a affaire à une indétermination (ceux qui seront plus « tolérants »). C'est bien ce que l'on observe, dans bien des cas, quand on travaille sur un corpus, y compris entre linguistes !

5.2. La méthode de calcul

Pour implémenter le modèle théorique que nous venons de présenter, il faut donc trouver un moyen de calculer la fonction potentielle associée à un co-texte donné. La méthode que notre équipe a conçue repose sur l'utilisation d'un très gros corpus : la fonction potentielle est obtenue à partir des données de cooccurrence des différents synonymes de l'unité étudiée avec le co-texte en question. Nous allons d'abord exposer l'idée générale sur l'exemple de *monter* dans le co-texte *escalier*, avant de présenter les détails techniques.

Il faut noter que cette méthode, au départ, ne traite que l'influence d'un seul élément de co-texte. Or, comme on l'a vu au chapitre 2 (§2.3. Polysémie verbale), pour désambiguïser un verbe, on a besoin de plusieurs éléments co-textuels. Nous verrons dans les prochains chapitres comment nous avons essayé de résoudre ce problème.

5.2.1 Le principe : calculer un degré d'affinité entre une unité co-textuelle et une clique

Pour calculer l'influence du co-texte *escalier* sur le sens de *monter*, on commence donc par relever les cooccurrences de *escalier* avec les 107 synonymes de *monter* dans un très gros corpus (un corpus de plusieurs centaines de millions de mots :

on verra plus bas le détail des corpus que nous avons utilisés). A partir de ces données, on va calculer un *degré d'affinité* de *escalier* avec chacun des synonymes de *monter.* L'hypothèse de base, c'est que l'on aura un degré d'affinité plus élevé pour des synonymes comme *grimper*, *gravir*, *se hisser*, qui correspondent au sens correct de *monter* avec ce co-texte que pour des synonymes qui ont d'autres sens, comme *augmenter*, *organiser* ou *préparer*. En fait cette hypothèse est loin d'être parfaitement exacte. D'abord, il existe des synonymes de *monter* qui ont le sens désiré et que l'on ne va sans doute pas trouver avec *escalier*, comme par exemple *ascensionner*. Et à l'inverse, il existe des synonymes que l'on va trouver avec *escalier* alors qu'ils ont un autre sens : ainsi, *fabriquer* est un synonyme de *monter*, et l'on peut certainement trouver dans un gros corpus *fabriquer un escalier* qui n'a pas le sens souhaité[39]. Il y a donc du « bruit » et du « silence » qui sont introduits par cette hypothèse. Mais cela n'est pas très grave, parce que notre calcul ne s'arrête pas là. En effet, ce que l'on va faire à partir de là, c'est travailler sur les cliques, et non sur les synonymes : on va donc calculer un degré d'affinité de *escalier* avec toutes les cliques de *monter*, en utilisant une pondération adéquate des degrés d'affinité avec les synonymes. Ainsi une clique comme {*ascensionner, escalader, gravir*, *grimper*, *monter*}, bien que contenant *ascensionner*, va pouvoir obtenir un très bon « score » et être sélectionnée comme un sens correct de *monter* avec *escalier*, tandis que la clique {*créer, fabriquer, faire, monter, réaliser*} aura un score beaucoup plus faible. De plus, ce sont toutes les cliques de la région de l'espace sémantique correspondant au sens d'élévation qui vont obtenir de bons scores, contrairement à la plupart des cliques de la zone correspondant au sens d'assemblage. On pourra donc obtenir une fonction potentielle correcte, même si l'hypothèse de départ n'est pas complètement fiable.

Notre méthode consiste donc en un calcul en trois étapes :

---

[39] A vrai dire, on pourrait théoriquement envisager que *monter un escalier* veuille dire effectivement *fabriquer un escalier*, mais en pratique le cas ne se présentera pas (sinon comme jeu de mots chez un Raymond Devos…).

- calcul d'un degré d'affinité du co-texte avec tous les synonymes de l'unité étudiée.

- calcul d'un degré d'affinité du co-texte avec toutes les cliques de l'unité étudiée.

- calcul de la fonction potentielle du co-texte sur l'espace sémantique de l'unité étudiée.

5.2.2 Calcul du degré d'affinité : première méthode

La première méthode mise au point par notre équipe consistait à pondérer les fréquences des cooccurrences par une méthode qui s'inspirait du calcul du χ^2, pour obtenir un degré d'affinité compris entre 0 et 1.

Plus précisément, appelons w_1, w_2, …, w_m les éléments du co-texte sur lesquels on veut travailler, et reprenons les notations du chapitre précédent : u_1, u_2, …, u_n désignent les synonymes, c_1, c_2, …, c_p les cliques, et $x_{ki} = 1$ ssi $u_i \in c_k$. Notons n_{ij} le nombre d'occurrences du couple (w_i , u_j) dans le corpus. Ce nombre doit être pondéré par la plus ou moins grande fréquence de w_i et de u_j, pris indépendamment, dans le corpus. S'il n'y avait pas d'affinité particulière entre certains éléments de co-texte et certains synonymes, les couples seraient équidistribués : autrement dit, le nombre d'occurrences du couple (w_i , u_j) ne devrait être fonction que de la fréquence des deux mots pris indépendamment. Appelons m_{ij} ce nombre moyen « théorique ». On peut montrer facilement que l'on a :

$$m_{ij} = \frac{n_{i\bullet}.n_{\bullet j}}{n},$$

avec des conventions analogues à celles du chapitre précédent :

$n_{i\bullet}$ est le nombre d'occurrences de w_i,

$n_{\bullet j}$ est le nombre d'occurrences de u_j,

et n est le nombre total d'occurrences de couples de type (w, u).

Pour mesurer l'affinité d'un élément de co-texte et d'un synonyme, il faut donc comparer n_{ij} et m_{ij}. Si n_{ij} est nettement supérieur à m_{ij}, cela veut dire qu'ils entretiennent

une relation d'affinité particulière. Si au contraire n_{ij} est nul ou nettement inférieur à m_{ij}, cela signifie que le couple est non attesté ou très rare.

Nous définissons donc le degré d'affinité d_{ij} de l'élément de co-texte w_i avec le synonyme u_j de la manière suivante :

$d_{ij} = f\left(\frac{m_{ij}}{n_{ij}}\right)$ où f est définie de la manière suivante :

$f(x) = \frac{x}{2}$ si $0 < x < 2$ et $f(x) = 1$ si $x > 2$

(le degré d'affinité est donc toujours compris entre 0 et 1).

Pour calculer le degré d'affinité avec une clique, on fait alors la somme pondérée des affinités avec toutes les unités qui constituent la clique. Plus précisément, le degré d'affinité a_{ik} de l'élément de co-texte w_i avec la clique c_k est donné par la formule suivante :

$a_{ik} = \frac{\sum_j d_{ij}.p_{ij}.x_{kj}}{\sum_j p_{ij}.x_{kj}}$ où le facteur de pondération p_{ij} vaut $\frac{m_{ij}}{\sum_k x_{kj}}$

5.2.3 Calcul du degré d'affinité : deuxième méthode

Lors des premières expérimentations que nous avons menées avec les verbes, nous avons ressenti le besoin d'améliorer ce calcul du degré d'affinité, parce qu'il n'était pas bien adapté à ce dont nous avions besoin. En effet, il nécessite de prendre en considération l'ensemble des éléments co-textuels pouvant se trouver à la place de l'élément que l'on étudie, et cela n'était pas très satisfaisant (il y avait une part d'arbitraire dans la sélection de cet ensemble) ni très pratique (notamment pour étudier l'influence des constructions syntaxiques).

Nous avons donc été amené à concevoir une autre méthode, basée sur le calcul de l'information mutuelle (Church et Hanks, 1990), dont on a vu (chapitre 3) qu'elle est

souvent utilisée dans les systèmes de désambiguïsation. Cette méthode ayant donné tout de suite de bien meilleurs résultats (cf. Jacquet, 2004), nous l'avons adoptée définitivement par la suite.

Appelons y_{ij} le quotient qui sert à calculer l'information mutuelle du couple (w_i , u_j).

Avec les notations de la section précédente, on a :

$$y_{ij} = \frac{\frac{n_{ij}}{N}}{\frac{n_{i\bullet} \times n_{\bullet j}}{N^2}}$$

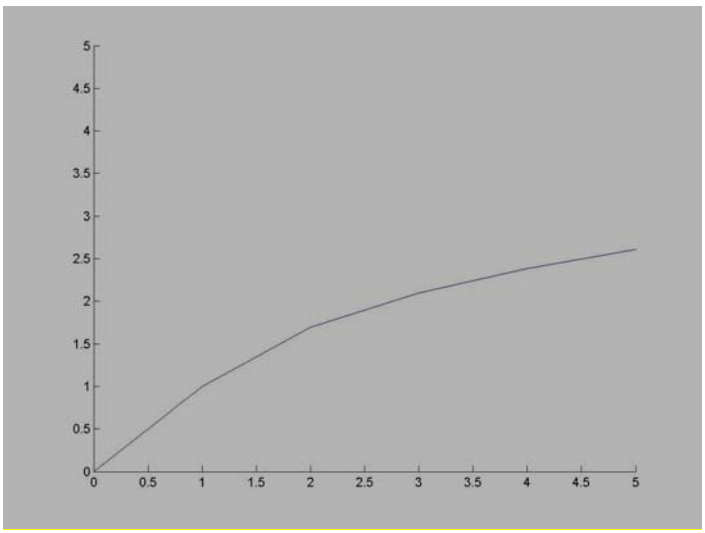
où N est le nombre de phrases du corpus

Plutôt que de prendre l'information mutuelle proprement dite, à savoir $\log(y_{ij})$, nous définissons le degré d'affinité de la manière suivante :

$d_{ij} = f(y_{ij})$ où f est définie de la manière suivante :

$f(y) = y$ si $0 < y < 1$ et $f(y) = \log(y) + 1$ si $y > 1$

Ci-dessous, le graphe de la fonction f.



Pour calculer le degré d'affinité avec une clique, on fait simplement la moyenne des affinités avec toutes les unités qui constituent la clique. Le degré d'affinité a_{ik} de l'élément de co-texte w_i avec la clique c_k est donc donné par la formule suivante :

$$a_{ik} = \frac{\sum_j d_{ij} . x_{kj}}{\sum_j x_{kj}}$$

5.2.4 Calcul de la fonction potentielle

Le calcul de la fonction potentielle associée à un élément de co-texte se fait par la même méthode que celle qui a été présentée au chapitre précédent (§ 4.3.4) pour les fonctions de pertinence d'un synonyme.

Autrement dit, la valeur de la fonction potentielle associée à l'élément de co-texte w_j au point de coordonnées $(x;y)$ est donnée par :

$$f_j(x,y) = \max\left(0, \sum_{k=1}^{c} a_{jk} \times e^{-\frac{(x_k - x)^2 + (y_k - y)^2}{\delta^2}}\right)$$

où (x_k,y_k) sont les coordonnées du point représentant la clique c_k dans l'espace sémantique,

a_{jk} désignant le degré d'affinité de c_k avec w_j

et $\delta = \frac{\max(dx,dy)}{10}$ avec $dx = xsup - xinf$ et $dy = ysup - yinf$

$$xsup = \max_{k=1}^{c}(x_i) \text{ et } xinf = \min_{k=1}^{c}(x_i)$$

$$ysup = \max_{k=1}^{c}(y_i) \text{ et } yinf = \min_{k=1}^{c}(y_i)$$

On verra de nombreux exemples de fonction potentielle calculée par cette méthode dans les chapitres suivants.

5.3. Les corpus

Les modes d'utilisation des corpus comme ressource en TAL sont considérables. Que ce soit pour la désambiguïsation comme nous l'avons vu, ou pour l'extraction de connaissances lexicographiques, la construction de résumé automatique et bien d'autres encore. Habert *et al.* (1997) proposent une description très complète de la construction et de l'utilisation des corpus en linguistique. Reprenons quelques points de la typologie de corpus présenté dans cet ouvrage.

Ils proposent de reprendre l'opposition entre *corpus* et *collection de texte* donnée par Sinclair (1996 : 4) pour définir ce qu'ils appellent un corpus : « un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage. » Cette définition des corpus est assez restrictive et se distingue des *collections de textes* qui « renvoient à des ensembles de textes [ne nécessitant pas] de sélection ou d'organisation, ou dont la sélection ou l'organisation ne nécessitent pas de critères linguistiques. »

Ils distinguent ensuite les *corpus de référence* des *corpus spécialisés* :

- Les **corpus de référence** sont conçus « pour fournir une information en profondeur sur une langue. Il[s] vise[nt] à être suffisamment étendu[s] pour représenter toutes les variétés pertinentes du langage et son vocabulaire caractéristique, de manière à pouvoir servir de base à des grammaires, des dictionnaires et d'autres usuels fiables. » (Habert *et al.*, 1997 : 144).
- **« Les corpus spécialisés** sont limités à une situation de communication, ou à un domaine. Parmi ces corpus on trouve les ensembles relevant de sous-langages que l'on trouve dans les domaines scientifiques et techniques. » (Habert *et al.*, 1997 : 144).

Nous avons précisé en introduction que notre objectif était de travailler sur la langue générale. Il semble qu'un corpus de référence soit le mieux adapté à notre étude. Cependant la réalité empirique nous montre qu'un tel corpus relève plus d'un idéal théorique que d'une ressource existante, du moins aujourd'hui pour le français.

Nous avons alors choisi deux bases textuelles qui selon nous étaient les plus proches de cet idéal : d'une part *Frantext* et d'autre part un regroupement d'articles tirés du journal *Le Monde*. Ces bases sont encore assez éloignées du corpus de référence idéal puisqu'elles correspondent à des textes littéraires pour *Frantext*, journalistiques pour *Le Monde*. De plus, nous allons voir que chacune regroupe des textes relevant de discours parfois éloignés. Nous avons donc plutôt à faire à des « collections de textes » (Habert *et al.*, 1997). Nous sommes conscient de ces lacunes, mais il n'en reste pas moins que ces deux bases de textes sont les plus proches du corpus référentiel idéal et pour cette raison nous leur donnerons dans cette étude le statut de « corpus »[40]. Nous proposons maintenant de décrire plus en détail ces deux corpus qui, comme nous allons le voir, se décomposent chacun en deux sous-corpus.

5.3.1 Frantext

Frantext (disponible sur le site Web http://www.atilf.fr/) peut être défini « comme l'association d'une part d'un vaste corpus de textes littéraires français, et d'autre part d'un logiciel offrant une interface Web avec des possibilités d'interrogation, de consultation et d'hyper-navigation. Historiquement, le but premier de ce corpus textuel était de permettre la constitution d'une base d'exemples destinée aux rédacteurs des articles du TLF (Trésor de la Langue Française). »

[40] Bourigault et Frérot (2005) disent en parlant du corpus du journal *Le Monde* : « Nous ne prétendons pas que ce corpus soit représentatif de la 'langue générale', mais nous considérons que sa taille et sa diversité thématique en font un corpus référentiellement et linguistiquement peu marqué. »

Cet outil contient 3 737 textes appartenant aux domaines des sciences, des arts, de la littérature, des techniques, qui couvrent 5 siècles de littérature (du XVIe au XXe siècle). Deux versions de Frantext sont proposées :

- l'intégralité de la base (3 737 textes, environ 210 millions d'occurrences, environ un millier d'auteurs). Les œuvres se répartissent en 80% d'œuvres littéraires et 20% d'œuvres scientifiques ou techniques.

- une sous-partie constituée de 1 940 œuvres en prose des XIXe et XXe siècles, soit environ 127 millions d'occurrences, qui ont fait l'objet d'un codage grammatical selon les parties du discours. Aux fonctionnalités du Frantext intégral, ont été ajoutées des possibilités de requêtes portant sur les codes grammaticaux.

C'est uniquement cette sous-partie, appelée *Frantext catégorisé*, que nous utiliserons dans cette étude. Pour simplifier la lecture, nous le nommerons régulièrement le *corpus Frantext*, mais il s'agit bien à chaque fois du corpus catégorisé.

5.3.2 Le journal Le Monde

Le corpus du journal *Le Monde* se décompose aussi en deux versions :

- le corpus LM10 : un corpus de 200 millions de mots, constitué des articles du journal *Le Monde*, des années 1991 à 2000. Ce corpus a été préparé, à partir de fichiers obtenus auprès de l'agence ELRA, par Benoît Habert (LIMSI), qui a effectué les tâches de nettoyage, de balisage et de signalisation nécessaires pour transformer les fichiers initiaux en un corpus effectivement « traitable » par des outils de Traitement Automatique des Langues. Pour cette étude, nous n'avons pas eu accès à l'ensemble du corpus mais simplement à l'ensemble des fréquences de triplets {recteur;relation;régi} dont nous parlerons dans le chapitre 9 (§9.1.).
- le corpus LM3 est un extrait du précédent (1994 à 1996 soit 11 millions de mots) avec une analyse syntaxique complète grâce à l'analyseur Syntex développé par D. Bourigault (ERSS). Cet analyseur s'appuie sur des relations de dépendance et sera décrit dans le chapitre 8.

6. Les constructions verbales : calcul de l'influence du co-texte lexical

L'objectif ici est de présenter le calcul de l'influence du lexique sur le sens d'un verbe. Dans notre modèle cela revient à calculer dans l'espace sémantique du verbe étudié, les régions de sens activées par le co-texte lexical, c'est-à-dire à calculer la fonction potentielle d'une unité co-textuelle lexicale sur l'espace sémantique du verbe étudié. Plusieurs études basées sur la construction dynamique du sens ont déjà été faites, notamment sur différents adjectifs dont *sec, curieux, gros, gras* (Manguin *et al.* 2002 ; François & al, 2003; Venant, 2004 ; François *et al.* 2005). Le caractère nouveau de cette partie sur l'étude du co-texte lexical est d'une part le fait d'appliquer la méthode aux verbes, et d'autre part l'implémentation d'un système logiciel permettant d'appliquer la méthode à n'importe quel mot et n'importe quel co-texte lexical de manière automatique, ce qui n'était pas le cas jusqu'à présent. Dans cette partie, nous présenterons quelques éléments de l'implémentation logicielle construite. Nous présenterons ensuite les résultats que nous avons obtenus puis discuterons de la qualité de ces résultats et des questions qui en découlent.

6.1. Logiciel

Le logiciel développé pour ce modèle pourrait être représenté par trois parties : 1- Calcul et affichage de l'espace sémantique de l'unité étudiée ; 2 - Calcul des degrés d'affinité pour un co-texte donné ; 3 - Calcul et affichage de la fonction potentielle du co-texte sur l'espace sémantique de l'unité étudiée.

Les parties 1 et 3 étaient déjà implémentées au début de cette étude. La partie 2 posait problème pour deux raisons. La principale raison était qu'initialement, les

fonctions potentielles étaient encore basées sur le calcul du degré d'affinité *première méthode* (cf. § 5.2.2). Comme nous l'avons vu, ces degrés d'affinité impliquent des calculs importants puisque pour avoir la fonction potentielle d'un co-texte sur un verbe, il faut calculer les fréquences relatives à tous les co-textes de ce verbe et calculer le degré d'affinité de chaque co-texte pour tous les synonymes de ce verbe. Ces calculs ne sont pas contestables du point de vue théorique, mais ils sont beaucoup trop coûteux en temps.

La deuxième raison était que quel que soit le mode de calcul du degré d'affinité, cela nécessitait des calculs de fréquence à partir du corpus Frantext. Or, l'interface Web de Frantext permet entre autre de rapatrier la fréquence d'un mot ou d'une expression, mais il n'était pas possible de rendre cette étape automatique.

Avec la collaboration de Jean-Marie Pierrel et de Jacques Dendien, nous avons fait en sorte que cette étape puisse devenir automatique. Le principe consiste à mettre en place un appel à distance du serveur de l'ATILF directement à partir de notre logiciel Visusyn programmé en Matlab. De cette manière on a la possibilité d'envoyer la liste des mots ou expressions qui nous intéressent, et le serveur nous retourne les fréquences respectives. Nous n'allons pas présenter l'intégralité des lignes de programmation nécessaires à cette étape (environ 400 lignes de code). Simplement, nous pointons sur deux moments critiques qui sont la traduction du langage courant de l'utilisateur en requêtes compréhensibles par le serveur ainsi que les commandes d'appel avec le serveur de l'ATILF.

L'utilisateur doit entrer la liste des mots ou expressions co-occurrents qu'il veut étudier à l'aide de la fenêtre de dialogue présentée figure 6.1 (le choix du mot à désambiguïser est fait avant). Il entre un mot ou une expression en tant que "graphie" et précise si besoin le type de graphie qu'il attend (Substantif, adjectif, verbe fléchi, etc.). Les expressions choisies s'affichent dans la partie droite de la fenêtre, déjà mises sous la forme de requêtes Frantext. Dans la figure 6.1 il y a trois co-textes enregistrés, le premier correspond au mot *marché* sans aucune autre information, le second correspond

à *marché* en tant que substantif fléchi (@m devant *marché* implique la flexion et g=cS impose la catégorie grammaticale substantif), et le troisième correspond à *marché* au singulier en tant que substantif précédé de *un*. La simple mise en forme de ces requêtes nécessite déjà 40 lignes de code (cf. annexe B.1.).

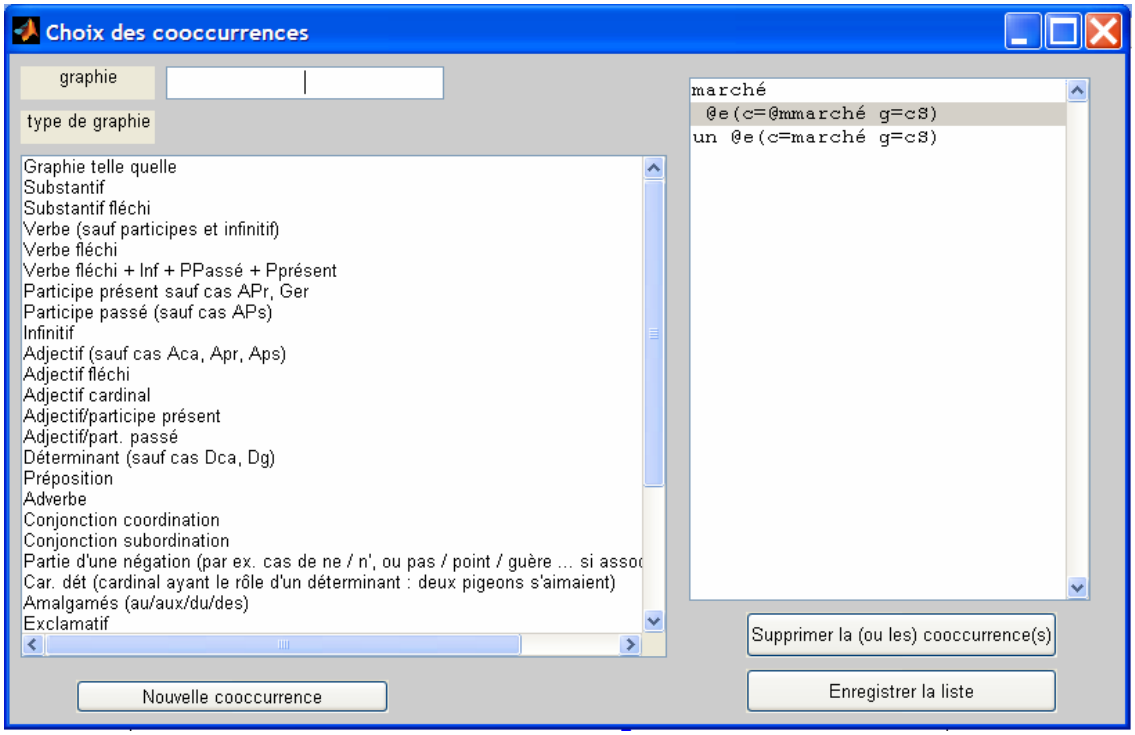


Figure 6.1 : Interface pour les requêtes Frantext

La deuxième étape importante est la construction de la requête d'appel pour extraire les fréquences à calculer dans Frantext via le serveur de l'ATILF La figure 6.2 présente la structure algorithmique de la fonction *requeteFrantext* dont le rôle est de construire la requête. On peut noter d'une part le nombre important de calcul de fréquences à effectuer puisque pour chaque couple (mot étudié, mot co-textuel), il faut calculer la fréquence du mot co-textuel dans le corpus, la fréquence de tous les synonymes du mot étudié dans le corpus (y compris le mot étudié) ainsi que la fréquence de tous les synonymes du mot étudié lorsqu'ils sont dans le même énoncé que le mot co-textuel. C'est pour cette raison que nous avons créé le fichier *fichierStockage* afin de stocker tous les calculs de fréquence déjà effectués. La figure 6.3 présente la

fonction *lancerRequeteFrantext* appelée à la fin de la fonction *requeteFrantext*, et qui correspond à l'envoi de la requête et à la réception des résultats à proprement parler.

```
requeteFrantext (listeSyn,listeCooc,fichierStockage)
        % listeSyn = la liste des synonymes du mot étudié
        % listeCooc = la liste des cooccurrences à étudier
        % fichierStockage = fichier de stockage des calculs de fréquence

        si fichierStockage existe alors l'ouvrir, sinon le créer
        requete = "";
        pour chaque cooccurrence i appartenant à listeCooc
            si freqCooc(i) n'existe pas dans fichierStockage
                requete = requete + requête de calcul de la fréquence de listeCooc(i)
            fin si
            pour chaque synonyme j appartenant à listeSyn
                si freqSyn(j) n'existe pas dans fichierStockage
                    requete = requete + requête de calcul de la fréquence de listeSyn(j)
                fin si
                si freqSynCooc(i,j) n'existe pas dans fichierStockage
                    requete = requete + requête de calcul de la fréquence de
                    listeCooc(i) + listeSyn(j) dans le même énoncé
                fin si
            fin pour
        fin pour
        exécuter lancerRequeteFrantext(requete)
        stocker les frequences calculées dans freqSyn, freqCooc et freqSynCooc
        stocker freqSyn, freqCooc et freqSynCooc dans fichierStockage
```

***Figure 6.2 : Structure algorithmique de la fonction* requeteFrantext**

```
function resultat = lancerRequeteFrantext(requete)
% fonction : envoyer une requête au serveur et de stocker les résultats renvoyés
% Le paramètre "requete" correspond à la requête en format texte
% retourne la variable "resultat" contenant les fréquences sous forme de vecteur Matlab

    %choix des noms de fichiers
    %respectivement pour les données d'entrée et de sortie
    nomFichier = 'c:\\wims\\www\\Frantext\\requeteFrantext.xml';
    donnee = ' c:\wims\www\Frantext\donneeRequete.xml';
    %stockage de la requête "requete" dans le fichier "requeteFrantext.xml"
    fid = fopen(nomFichier, 'w');
    fwrite(fid, requete);
    fclose(fid);
    %envoie de la requête au serveur de l'ATILF et réception des résultats
    dos(strcat('C:\Frantext\send.exe',...
               ' http://stella.atilf.fr/dendien/scripts/special/gest.exe',...
               ' c:\wims\www\Frantext\requeteFrantext.xml',...
              donnee));
    %extraction des fréquences sous forme de vecteur Matlab
    %à partir du fichier de résultat
    %à l'aide du programme perl extraireFreq.pl
    [status,resultat] = dos(strcat('C:\perl\bin\perl',...
                                   ' c:\wims\www\Frantext\extraireFreq.pl',...
                                   donnee));
```

***Figure 6.3 : Fonction* lancerRequeteFrantext**

6.2. Résultats : Région sémantique associée à un nom co-textuel

Nous proposons, pour l'instant, de poursuivre notre présentation avec le verbe *monter*. Quelques résultats sur l'influence du co-texte lexical sur le verbe *jouer* seront présentés dans le chapitre 8 (§ 8.5). Rappelons les principes du calcul : soit un mot à désambiguïser *w*, et un co-texte lexical *c*. On extrait du corpus les fréquences relatives de *c* lorsqu'il cooccurre avec chaque synonyme s_i de *w*, dont *w*. Dans ce cas, le corpus est *Frantext catégorisé*, et les fréquences sont extraites via la fonction de requête à distance présentée dans le paragraphe précédent. La fonction potentielle de *c* sur l'espace sémantique de *w* est calculée à partir de l'ensemble des degrés d'affinité[41] $a(u_i,c)$.

Nous proposons d'illustrer les résultats obtenus à l'aide de trois co-textes lexicaux : *escalier*, *diamant* et *projet*. *Escalier* est censé illustrer le fonctionnement du modèle avec un mot fréquemment employé avec le verbe *monter*. *Diamant* illustre le cas de mots très peu employés avec *monter*, et correspondant à un sens très réduit de *monter*. Enfin, *projet* permet de mettre en évidence certaines erreurs du modèle (cf. figure 6.4 pour un récapitulatif des fréquences). Pour ces trois co-textes, notre description sera basée d'une part sur la région calculée par notre modèle et d'autre part sur les cliques dont le degré d'affinité est supérieur ou égal à un, c'est-à-dire les degrés d'affinité les plus élevés.

Pour tous les co-textes lexicaux étudiés, l'idée est de s'intéresser aux compléments du verbe, plus qu'au sujet. Avec le postulat que le complément d'un verbe est plus déterminant pour sa désambiguïsation que son sujet. Audibert (2003) montre que les mots les plus désambiguïsants pour un verbe *v* sont ceux qui sont en première, deuxième et troisième position après *v*. Les calculs de fréquence sont faits avec la contrainte suivante : le co-texte doit se trouver après le verbe *v*, *v* étant un des

[41] Tous les degrés d'affinité sont calculés avec la deuxième méthode, basée sur l'information mutuelle (cf.§ 5.2.3).

synonymes du verbe vedette, et il ne doit pas y avoir de verbe entre *v* et le co-texte. Ceci afin d'accentuer les chances que le co-texte ait une quelconque relation avec *v*.

Frantext catégorisé	Total	co-texte *escalier*	co-texte *diamant*	co-texte *projet*
Fréquence avec monter	24 917	980	4	0
Fréquence totale	-----	23 107	3 567	12 197

Figure 6.4 : Récapitulatif des fréquences dans Frantext

6.2.1 Monter *et le co-texte* escalier

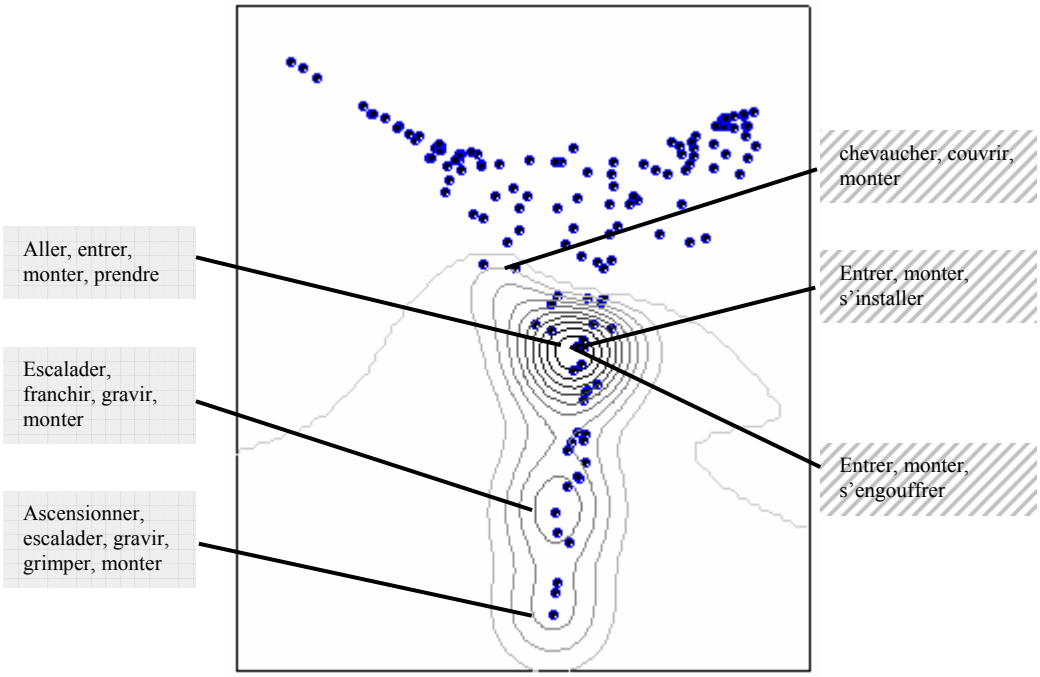


***Figure 6.5 : Fonction potentielle du co-texte* escalier**

La définition de *monter* dans le TLFi cite deux fois le co-texte *escalier* :

I.A.3. Parcourir en s'élevant, grimper, faire l'escalade de. *Monter un escalier*

II.A.1.S'élever pour se trouver plus haut. Monter par un escalier/ monter les degrés d'un escalier

La région activée par notre modèle (cf. figure 6.5) correspond à l'ensemble de l'axe 1 que nous avions nommé « être vivant se déplaçant dans un mouvement ascendant/se rendant dans un endroit plus haut que là où il se trouve ». On peut observer

trois minima locaux au sein de la région activée. Ils correspondent de haut en bas aux cliques :

Aller, entrer, monter, prendre

Escalader, franchir, gravir, monter

Ascensionner, escalader, gravir, grimper, monter

Regardons maintenant les cliques qui ont le plus contribué à la construction de cette région de sens :

N° de clique	**cliques**	**Degré d'affinité**
1	*escalader, franchir, gravir, monter*	1
2	*entrer, monter, s'engouffrer*	1
3	*ascensionner, escalader, gravir, grimper, monter*	1
4	*monter, s'envoler, s'échapper*	1
5	*chevaucher, couvrir, monter*	1
6	*entrer, monter, s'installer*	1
7	*atteindre, gravir, monter*	1
8	*escalader, grimper, monter, se hisser*	0,8
9	*assembler, bâtir, coudre, monter*	0,7

Notons que les cliques 1, 3, 4 et 7 correspondent aux sens attendus. La clique 6 (*entrer, monter, s'installer)* est discutable puisque si le verbe *entrer* convient (*monter/entrer par l'escalier*), le degré d'affinité entre *s'installer* et *escalier* provient d'un seul énoncé du type *s'installer en haut de l'escalier* mais dans ce cas, *s'installer* ne correspond pas au sens de *monter*.

> « Les trois plus grands d'entre ces géants recrutés traditionnellement au Soudan, **s'installaient** en haut de l'**escalier** de marbre, sur le palier du cabinet de travail de Ben Akbir. » (Déon. M, La carotte et le bâton, 1960, page 268).

On retrouve le même cas de figure avec la clique 2 et le verbe *s'engouffrer (s'engouffrer dans l'escalier* qui peut être employé aussi bien pour *monter* que pour *descendre les escaliers)*.

> « des hommes en blouse blanche, appuyés sur leurs fusils, sont montés sur les piédestaux des colonnes du péristyle et crient : entrée libre du bazar, pendant que la foule fait irruption, les chapeaux en l' air, et qu' une immense clameur s' **engouffre** dans l' **escalier** du palais envahi. » (Goncourt E. et J. , Journal T.2, 1878, page 590)

Quand à la clique 5 (*chevaucher, couvrir, monter*), elle correspond à une erreur. Son degré d'affinité est élevé par la présence de quatre énoncés contenant *couvrir* et *escalier* mais qui ne correspondent pas à un rattachement au verbe *couvrir* d'un complément contenant *escalier,* ainsi que par la présence de deux énoncés contenant *chevaucher* et *escalier* pour lesquels nous ne nous aventurerons pas dans une interprétation…

> « La princesse s'arrêta d'un air moqueur, et le bruit de sa robe de soie, épaisse et cassante comme du carton, cessant de **couvrir** les bruits plus éloignés, nos trois héroïnes, parvenues presque à la grande cage d'**escalier** qui s'ouvrait au fond de la galerie, entendirent distinctement le bruit sec d'un balai » (Sand G., La comtesse de Rudolstadt, 1843, page 97)

> « Une voilette **couvrait** à demi son visage très pâle, et sur sa poitrine, agitée par l'émotion et par la montée de l'**escalier**, elle serrait nerveusement un objet enveloppé dans un journal. » (Theuriet A., La maison des 2 Barbeaux, 1879, page 140)

> « D'ailleurs, non, je ne vois pas cela, ce piano, même démonté, **chevauchant** la nuit dans les **escaliers** chaotiques de l'Iran. » (Loti P., Vers Ispahan, 1904, page 919).

> « Elle a trois palais, plus riches, plus ornés que ceux de *Caceres[42], mais qui ont moins de grandeur, et une belle église, celle de *San *Martin, près de laquelle une statue équestre de *Pizarro, d'un bronze vert tout neuf, **chevauche** au haut d'un **escalier**. » (T'serstevens A., L'itinéraire Espagnol, 1963, page 199).

[42] Dans tous les extraits de Frantext que nous citerons, certains noms propres sont précédés du caractère *. Cela correspond à un balisage propre à Frantext qui n'est évidemment pas présent dans le texte initial.

Ce qu'il faut retenir, c'est le rôle crucial de la fonction potentielle. En effet, classer les cliques par ordre de degré d'affinité ne suffit pas à représenter les sens possibles pour un co-texte donné. C'est les interactions entre ces degrés d'affinité qui vont faire émerger les régions de sens pertinentes. Ainsi, on peut observer que la clique 5 est complètement exclue de la région activée, et les cliques 2 et 6 se trouvent dans la région activée puisqu'elle sont proches d'autres cliques plus correctes qui contiennent le verbe *entrer*, mais elles ne font qu'activer de manière exagérée la zone *entrer/monter par les escaliers* sans pour autant effacer les sens plus attendus (*grimper, gravir,* etc.)

6.2.2 Monter *et le co-texte* diamant

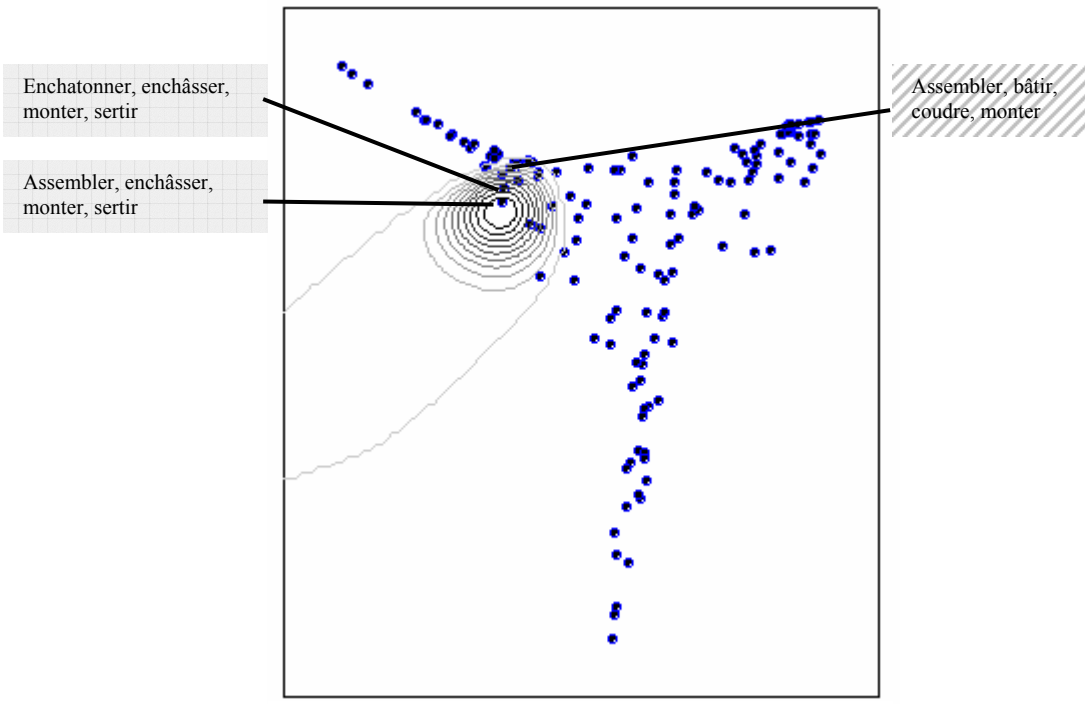


Figure 6.6 : Fonction potentielle du co-texte

L'intérêt du co-texte *diamant* est de montrer qu'il est possible d'appliquer notre calcul même lorsque la fréquence de la relation co-textuelle est faible : il n'y a que quatre occurrences de *monter* avec le co-texte *diamant* dans tout Frantext catégorisé. Etant donné ce faible nombre, nous nous permettons de présenter ces quatre énoncés :

> « - Oui, monsieur, mais les moyens d'exécution y sont pour quelque chose : je vous ai bien **monté** votre **diamant**. » (Balzac H. De, Histoire. de Cesar Birotteau, 1837, page 225)

> « Ils regardent un mari comme un ouvrier chargé de dégrossir, polir, tailler à facettes et **monter** le **diamant** qui passera de main en main, pour être un jour admiré à la ronde. » (Balzac H. De, Physiologie du mariage, 1846, page 989)

> « Au lieu que les belles images feront un style correct, pour la même raison qui fait qu'on ne **monte** pas un beau **diamant** sur cuivre. » **(**Alain, Propos, 1936, page 134)

> « Les arceaux chargés de roses, les grandes masses d'arbres noirs, les pelouses inclinées, la brume bleuâtre des lointains sous un ciel pâle où **montait** une lune de **diamant**, tout cela était d'une beauté de féerie en même temps que légèrement

banale, mais banale comme ces merveilleux décors de *Jusseaume pour *Pelléas, qui ravissaient mes vingt ans. » (Green J., Journal T.4, 1946, page 116)

Les trois premiers extraits correspondent à l'idée que l'on se fait de *monter un diamant*, c'est-à-dire *monter un diamant sur une bague, sur un bijou*. En revanche, dans l'extrait de Green (1946), *diamant* est rattaché à *lune* et non à *monter* (*montait une lune de diamant*). En réalité, les quatre énoncés pourraient ne pas correspondre au sens attendu (*monter un diamant sur une bague*) sans pour autant compromettre notre calcul. Ce sont les synonymes de *monter* qui vont permettre un calcul fiable, en l'occurrence ici les synonymes *sertir* et *enchâsser*. En effet, il y a dans le corpus 32 occurrences de *sertir* dont 3 avec le co-texte *diamant*, de même qu'il y a 48 occurrences de *enchâsser* dont 4 avec le co-texte *diamant*. Les fréquences relatives pour ces deux synonymes sont élevées (à titre de comparaison, il y a 24 917 occurrences de *monter*, dont 4 avec le co-texte *diamant*) et vont être les principaux contributeurs à la fonction potentielle représentée figure 6.6. On les retrouve évidemment parmi les cliques ayant un degré d'affinité égale à 1 (cliques 1 et 2 du tableau qui suit).

N° de clique	**cliques**	**Degré d'affinité**
1	*enchatonner, enchâsser, monter, sertir*	1
2	*assembler, enchâsser, monter, sertir*	1
3	*assembler, bâtir, coudre, monter*	1
4	*enlever, jouer, monter*	0,4
5	*couvrir, faire, monter, servir*	0,4

La clique 3 (*assembler, bâtir, coudre, monter*) est sujette à discussion. A première vue, elle semble correspondre à une erreur. En regardant de plus près, on constate que c'est le verbe *coudre* qui contribue quasiment exclusivement au score de la clique (assembler : 0 ; bâtir : 0,36 ; coudre : 3,74 ; monter : 0,01 ; clique total : $\frac{0+0,36+3,74+0,01}{4}=1,03$)

Or, le fort score de *coudre* n'est dû qu'à un énoncé sur les 1 151 occurrences de ce verbe. Cet énoncé correspond effectivement au sens de *coudre/monter un diamant sur un bijou,* mais il aurait très bien pu correspondre à une erreur.

> « seulement, au centre de chaque rosette qu'ils portaient sur l' oreille, le comte avait fait **coudre** un **diamant**. » (Dumas A. Père, Le Comte De Monte-Cristo, 1846, page 694)

Ce cas met en évidence le fait que le calcul de l'information mutuelle n'est fiable que pour des fréquences élevées. Pour réduire l'influence des occurrences à faible fréquence, sans pour autant les supprimer, il serait intéressant de pondérer les informations mutuelles calculées en attribuant une note de fiabilité à chaque calcul (cf. chapitre 10 § 10.1.1). Toujours est-il que, malgré le fort degré d'affinité de cette clique incertaine, l'effet de lissage de la fonction potentielle permet de la renvoyer à la périphérie de la région activée.

Le cas de *diamant* est aussi intéressant par la taille réduite de la région activée, qui correspond bien à un sens précis de *monter*. On retrouve ces caractéristiques dans la liste des cliques du tableau ci-dessus puisque le degré d'affinité le plus élevée après les cliques 1,2 et 3, est seulement de 0,4. Il y a donc une séparation nette entre le sens véhiculé par les trois premières cliques et les autres. Ce n'est pas le cas pour le co-texte *escalier* où les degrés d'affinité décroissent beaucoup plus progressivement.

6.2.3 Monter *et le co-texte* projet

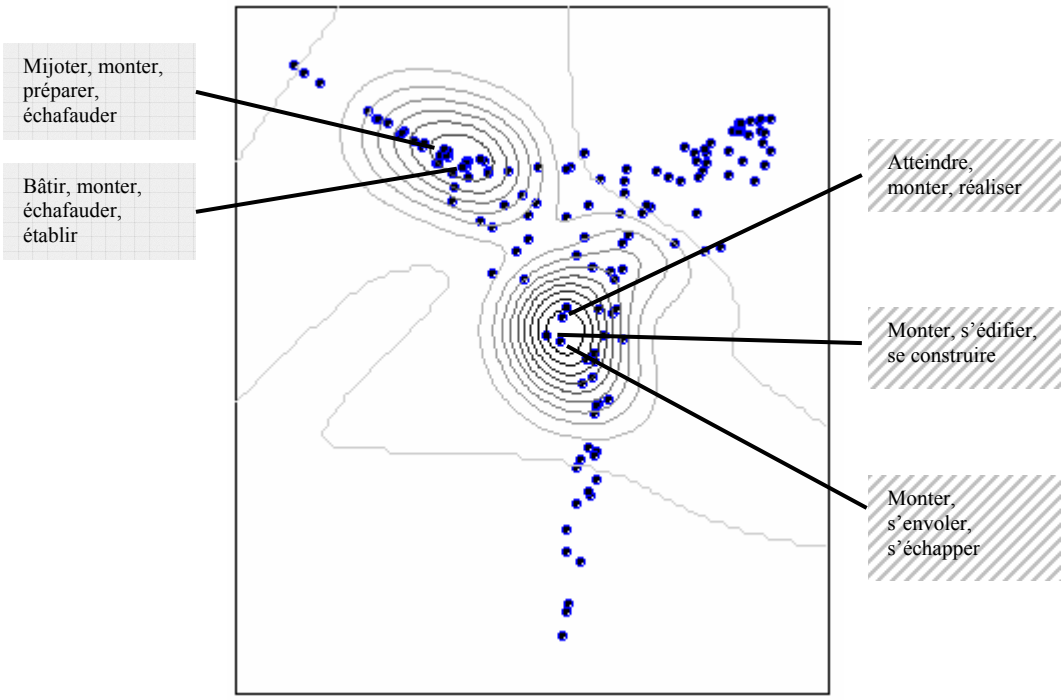


***Figure 6.7 : Fonction potentielle du co-texte* projet**

N° de clique	**cliques**	**Degré d'affinité**
1	monter, s'édifier, se construire	1
2	mijoter, monter, préparer, échafauder	1
3	atteindre, monter, réaliser	0,9
4	bâtir, monter, échafauder, établir	0,8

Le co-texte *projet* est intéressant pour deux choses. D'une part parce qu'il n'existe pas d'énoncé dans Frantext contenant *monter* et le co-texte *projet*. Cette absence est probablement due au caractère très littéraire du corpus Frantext. Il est fort probable que des énoncés contenant *monter un projet* seraient plus courants dans un corpus journalistique. Toujours est-il que notre modèle est capable de « spéculer » sur le sens d'un verbe en fonction de relations co-textuelles qui n'existent pas. En réalité, le

principe est le même que pour le co-texte *diamant*, c'est-à-dire que ce sont les synonymes de *monter* qui vont guider le comportement sémantique de *monter* avec ce co-texte.

D'autre part, ce co-texte illustre le type d'erreur que notre modèle peut faire. La figure 6.7 présente les deux régions activées par ce co-texte. L'une correspond à la zone de sens *mijoter, monter, préparer, échafauder* qui correspond au sens que l'on attendrait pour *monter un projet*, alors que l'autre correspond à la zone *monter, s'édifier, se construire*. Cette dernière zone est plus surprenante. Si l'on étudie les cliques responsables de l'activation de cette zone on retrouve (le numéro précédant chaque clique correspond au classement par ordre de degré d'affinité et la valeur du degré est mise entre parenthèses):

1 : *monter, s'édifier, se construire* (1)

3 : *atteindre, monter, réaliser* (0,9)

10 : *monter, s'envoler, s'échapper* (0,6)

La clique 3 (*Atteindre, monter, réaliser*) pourrait correspondre à un sens secondaire de *monter* introduisant la notion d'accomplissement. *Monter un projet* au sens *accomplir/réaliser un projet*. Toutefois, les deux autres cliques ne correspondent à aucun sens de *monter* lorsqu'il est en présence de *projet*. Pire, il n'y a aucune relation sémantique claire entre ces trois cliques, et pour cause, les cliques 1 et 10 n'obtiennent un tel score que par un synonyme, respectivement *se construire* et *s'échapper* et ces synonymes n'ont un tel degré d'affinité que par un énoncé.

> « De Coantré, tout à fait hors de sa norme, se **construisait** un **projet** sensationnel : celui de dîner au restaurant et d'aller ensuite à *Montmartre, à *Montmartre dont la basilique, par les rues transversales, apparaissait bizarrement proche, et comme à portée de la main. » (Montherlant H. De, Les célibataires, 1934, page 837).

> « au dire des autres prisonniers, il s'était **échappé** avec le **projet** d' aller aux *Indes. » (Balzac H. De, La peau de chagrin, 1831, page 106).

Autrement dit, les conditions pour qu'une région soit activée par erreur sont assez importantes puisque dans ce cas cela a nécessité deux cliques activées par erreur se trouvant dans la même partie de l'espace sémantique qu'une clique correspondant à un sens secondaire. Nous apporterons quelques remarques supplémentaires sur cette erreur dans la discussion qui suit.

6.3. Discussion

Les résultats obtenus sur les trois co-textes sont assez encourageants, puisque nous avons pu désambiguïser, ou du moins réduire l'ambiguïté d'un verbe simplement à l'aide d'une unité lexicale co-textuelle. Et ce en respectant le cadre du continu que nous nous sommes fixé. Notons aussi que notre modèle est assez résistant aux erreurs de relations co-textuelles calculées en amont puisque, comme nous l'avons vu pour *escalier* et *diamant*, certaines cliques ayant un degré d'affinité élevé par erreur n'impliquent pas nécessairement l'activation d'une région de l'espace. Il arrive malgré tout que certaines zones de sens soient activées par erreur ou suractivées, comme c'est le cas pour *projet*. Notons que dans le cas du co-texte *projet*, il est difficile de déterminer si la zone correspondant au sens « *atteindre, monter, réaliser* » méritait d'être activée. L'important pour notre modèle est de privilégier la « suractivation ». En effet, il est préférable d'être moins exigeant sur la réduction de l'ambiguïté que de supprimer à tort des sens possibles.

Indirectement, cette erreur a soulevé une autre question. Pourquoi des sens si distincts que *« monter, s'édifier, se construire »*, *« atteindre, monter, réaliser »* et *« monter, s'envoler, s'échapper »* se retrouvent-ils dans la même zone d'un espace sémantique censé refléter les proximités de sens ? Cela ne correspond pas à une erreur mais à un biais introduit par les dimensions extraites de l'AFC (Analyse Factorielle des Correspondances). Comme nous l'avons décrit dans le paragraphe 4.3.1, l'espace sémantique d'un mot *w* a autant de dimensions que *w* a de synonymes. Dans le cas de *monter*, 107 dimensions. Les deux dimensions présentées ici correspondent à celles qui

synthétisent le mieux l'information contenue dans les *n* dimensions, c'est-à-dire celles qui perdent le moins d'information. La conséquence étant que les cliques bien représentées par les deux meilleures dimensions auront des valeurs élevées sur les deux axes de ces dimensions, alors que les autres sens auront des valeurs proches de zéro. Concrètement, ces derniers se retrouveront au centre de l'espace, regroupant des sens divers mais qui n'interviennent pas dans les axes dominants.

En réalité, c'est ce biais qui est le principal responsable de l'erreur décrite pour le co-texte *projet*. Nous décrirons dans le chapitre 10 (§ 10.1.1) une solution à ce problème en basant le calcul de fonction potentielle non sur deux dimensions mais sur un nombre plus important de dimensions.

Le second point sur lequel nous souhaitons discuter concerne l'évaluation. Il est clair que l'étude de trois co-textes différents pour un même verbe ne suffit pas à évaluer un modèle. Ça n'était d'ailleurs pas l'objectif. L'idée était plutôt d'illustrer le fonctionnement de notre modèle par des cas intéressants de désambiguïsation, permettant du même coup de pointer sur les intérêts et les limites de ce modèle. Une évaluation a été faite sur un modèle similaire par Venant (2004). Dans cette étude, Venant évalue la capacité du modèle à désambiguïser l'adjectif *sec* en fonction de son nom recteur en prenant comme référence une étude psycholinguistique. La précision du modèle de désambiguïsation par rapport à l'étude psycholinguistique est de 79%, soit une précision très encourageante. Nous n'avons pas fait une telle évaluation pour les verbes, d'une part parce que l'influence du co-texte lexical n'est pas centrale dans notre étude, et d'autre part parce que faire ce type d'évaluation sur les verbes semble voué à l'échec. En effet, il est beaucoup plus difficile pour les verbes de dissocier l'influence du lexique de l'influence de la syntaxe. Ainsi, l'incompréhension d'un évaluateur face à un couple (*monter ; projet*) combiné à l'ambiguïté de la fonction potentielle de *projet* sur l'espace sémantique de *monter* aurait certainement fourni des taux d'adéquation très faibles.

7. Les Grammaires Constructionnelles

L'objectif de ce chapitre est de présenter le cadre théorique sur lequel nous nous sommes basé pour traiter les relations syntaxiques dans notre modèle de désambiguïsation. Nous avons fait le choix des grammaires constructionnelles et nous proposons ici de justifier ce choix. Nous commencerons par présenter les caractéristiques principales de cette théorie, développée entre autres par Fillmore, Kay et O'Connor (1988), Lakoff (1987), Lambrecht (1994), Fillmore et Kay (1996), que nous confronterons avec d'autres types d'approches telles que la *Role and Reference Grammar* de Van Valin (2004), le lexique-grammaire de Gross (1975, 1984) ou encore l'approche projectionniste de Levin (1993). Dans un deuxième temps nous développerons plus en détail une des variantes les plus connues de grammaire constructionnelle qui est l'approche d'Adèle Goldberg (1995). Nous terminerons notre description par quelques discussions et critiques des travaux de Goldberg.

7.1. Généralité sur les grammaires constructionnelles

Le point de départ des grammaires constructionnelles est de considérer les phrases d'une langue comme des instances de constructions, les constructions étant considérées comme des correspondances forme-sens qui existent indépendamment d'éléments lexicaux particuliers. Cela revient à considérer que les constructions sont porteuses d'un sens intrinsèque indépendamment des mots présents dans la phrase. Pour illustrer ce point, considérons les exemples suivants :

(1) Bees are swarming in the garden. (Il y a plein d'abeilles dans le jardin.)
(2) The garden is swarming with bees. (Le jardin est plein d'abeilles.)
(3) I am afraid to cross the road. (J'ai peur de traverser la route)
(4) I am afraid of crossing the road. (J'ai peur de la traversée de la route)

Il y a des différences de sens entre les énoncés (1) et (2) d'une part, et les énoncés (3) et (4) d'autre part. L'énoncé (2) implique qu'il y a des abeilles dans tout le jardin, alors que l'on peut énoncer (1) s'il n'y a des abeilles que dans une partie du jardin. De même, entre (3) et (4), seul l'énoncé (3) suppose une intention de traverser la route de la part du locuteur. Comme ce sont les mêmes unités lexicales qui sont présentes dans ces couples d'énoncés, ces différences de sens ne peuvent être attribuées qu'à la construction elle-même.

L'objectif des grammaires constructionnelles est donc de développer un modèle permettant de caractériser l'ensemble des constructions d'une langue, et pas seulement les structures qui sont définies comme appartenant à la grammaire. Elles visent à rendre compte des conditions dans lesquelles une construction donnée est considérée comme correcte ou non du point de vue grammatical. Autrement dit, comme bien d'autres théories linguistiques, elles tentent d'expliquer tant le fait qu'il existe un nombre infini d'expressions qui sont permises dans une langue que le fait qu'une infinité d'autres sont considérées comme incorrectes, mais, pour atteindre cet objectif, elles se centrent sur l'étude des constructions, plutôt que sur des mécanismes syntaxiques généraux, ou sur des données spécifiques lexicales.

Ce modèle est basé sur le fait qu'il n'y a pas de distinction stricte entre le lexique et la syntaxe : il y a un continuum entre des constructions très spécifiques de quelques éléments lexicaux et les constructions syntaxiques les plus génériques. Les grammaires constructionnelles rejettent aussi l'idée d'une division stricte entre la sémantique et la pragmatique. Les informations de type thématique (focus, thème, rhème, etc.) et de registre de langue sont représentées dans les constructions au côté des informations sémantiques. En cela elles s'opposent à la *Role and Reference Grammar* (RRG) de Van Valin (2004) qui propose des interfaces spécifiques entre syntaxe, sémantique et pragmatique. Autrement dit, là où la RRG construit des composantes distinctes (structure logique, projection des constituants, projection des opérateurs et

structure pragmatique) qui sont ensuite mises en relation, les grammaires constructionnelles posent d'emblée un cadre de représentation unique.

Parmi les approches qui peuvent *a priori* ressembler aux grammaires constructionnelles, on peut citer le lexique-grammaire, qui a été développé par Maurice Gross (1975, 1984). Cette théorie a en commun avec les grammaires constructionnelles la volonté de rendre compte des relations entre constructions grammaticales, notamment constructions verbales, et sens. Pour cela il recense pour chaque unité lexicale toutes les structures syntaxiques élémentaires et les sens associés. Une des idées importantes régissant le lexique-grammaire est de considérer que la séparation entre grammaire et lexique dans la description linguistique est contre-productive. A première vue cette caractéristique pourrait être un point commun avec la grammaire constructionnelle, c'est en réalité un point important d'opposition. Lorsque Gross dit que la séparation entre grammaire et lexique est contre-productive dans la description linguistique, il entend par là que grammaire et lexique sont deux parties bien distinctes de la description linguistique mais qu'il faut les traiter simultanément. A l'inverse la grammaire constructionnelle propose d'appréhender la correspondance forme-sens de la même manière pour la syntaxe et pour le lexique mais en distinguant en revanche l'apport de sens de ces deux éléments constituants de la phrase.

L'approche projectionniste de Levin (Levin, 1993, Rappaport Hovav et Levin 1998), a en commun avec l'approche constructionnelle la même appréhension du sens des phrases par l'étude du verbe et de ses arguments. Les deux approches divergent quant à la répartition de la part de sens attribuée à la syntaxe et celle attribuée au lexique. Levin considère que la sous-catégorisation syntaxique des verbes peut être prédéterminée uniquement à l'aide de la sémantique lexicale du verbe. En grammaire constructionnelle, constructions et verbes sont en interrelation, mais sont indépendants.

Par rapport à ces approches, les grammaires constructionnelles ont deux avantages importants. Le premier avantage est d'éviter l'impossibilité de traiter le sens d'un verbe. Ainsi, considérons les exemples suivants :

(5) He sneezed the napkin off the table *(Il a fait valser la serviette de la table en éternuant)*

(6) She baked him a cake *(Elle a fait cuire un gâteau pour lui)*

(7) Herman hammered the metal flat *(Herman a aplati le métal à coups de marteau)*

Dans chacun de ces énoncés, le verbe possède un argument supplémentaire, par rapport à son fonctionnement habituel, et cela entraîne l'apparition d'un sens qui n'est pas directement déductible de son sens habituel. Avec l'approche constructionnelle, il est possible de comprendre le mécanisme qui produit une interprétation finale de type 'mouvement induit' (*caused motion*) pour (5), 'transfert intentionnel' (*intended transfert*) pour (6) et 'résultat obtenu' (*caused result*) pour (7). Le sens global n'est pas construit à partir des unités lexicales directement, mais par l'intermédiaire des constructions auxquelles sont associées ces interprétations.

Le second avantage est de rendre compte de la polysémie des constructions. Levin avance comme argument le fait que la sémantique d'énoncés complets est différente à chaque fois que l'on change de verbe, ceci quelle que soit la construction. Autrement dit, une même construction peut changer de sens en fonction du verbe qu'elle contient. Mais ces différences n'ont pas besoin d'être nécessairement attribuées aux différents sens des verbes, elles peuvent aussi être attribuées aux constructions elles-mêmes. Autrement dit, si l'on admet que les constructions syntaxiques peuvent être polysémiques, au même titre que les unités lexicales, on peut rendre compte de manière beaucoup plus économique de cette diversité de sens.

7.2. La grammaire constructionnelle de Goldberg

Parmi les différentes formulations théoriques des grammaires constructionnelles, nous allons nous focaliser sur un modèle en particulier, celui d'Adèle Goldberg (1995). Nous avons choisi ce modèle pour deux raisons. D'une part parce que Goldberg décrit une partie de son modèle à l'aide d'un formalisme qui nous a semblé intéressant. La

deuxième raison est qu'elle oriente particulièrement son étude sur les structures argumentales de verbes, ce qui nous concerne directement. Nous allons d'abord développer la description des interactions entre verbes et constructions proposée par Goldberg. Puis, nous décrirons plus en détail la construction du 'mouvement induit' (*caused motion construction*).

7.2.1 Le modèle d'interaction entre verbes et constructions

L'idée maîtresse de la grammaire constructionnelle est donc de considérer que les constructions sont porteuses de sens indépendamment du lexique, et, plus particulièrement, du verbe. Il est clair que cela ne veut pas dire que les constructions imposent leur sens aux verbes. Les sens de la construction et du verbe interagissent de manière complexe. Cette double influence implique deux mécanismes simultanés d'analyse du sens. C'est ce que Goldberg propose de décrire.

La nature du sens des verbes et des constructions

D'un côté, les verbes possèdent un cadre sémantique dans lequel se définissent leurs sens. Dans l'esprit des grammaires cognitives, ce cadre sémantique est une structure d'arrière-plan complexe[43] incluant des connaissances sur le monde et des connaissances culturelles. Il doit permettre de capturer la richesse de tous les sens du verbe, et il est donc nécessaire pour son interprétation dans une construction donnée. Ce cadre sémantique contribue donc au sens global. C'est ainsi que la différence de sens entre les deux énoncés suivants provient de la différence de sémantisme entre les deux verbes.

---

[43] Goldberg fait explicitement référence à Lakoff (1987) qui défini les concepts par un ensemble de structures distinctes, ou modèles cognitifs idéalisés. Par exemple, pour le concept *mother* (mère), Lakoff soutient que cinq modèles cognitifs distincts sont à l'œuvre, qui ne se recouvrent pas forcément :

- le modèle de naissance (*birth model*) : la personne qui donne la vie à l'enfant.
- le modèle génétique (*genetic model*) : la femme qui apporte une contribution génétique à l'enfant.
- le modèle nourricier (*nurturance model*) : la femme qui nourrit et qui élève l'enfant.
- le modèle marital (*marital model*) : la femme du père.
- le modèle généalogique (*genealogic model*) : l'ancêtre femme la plus proche.

(8) Sally skipped over the crack in the ground *(Sally sauta par dessus la crevasse)*

(9) Sally crawled over the crack in the ground *(Sally rampa par dessus la crevasse)*

Notamment, si l'on peut déduire que Sally a été en contact avec la crevasse dans l'énoncé (9) et pas dans l'énoncé (8), c'est bien à cause du sémantisme spécifique de chacun des verbes utilisés.

De l'autre côté, les constructions sont elles aussi associées de manière typique à une famille de sens liés entre eux, plutôt qu'à un unique sens fixé et abstrait. C'est-à-dire qu'à partir du moment où les constructions sont traitées de la même manière que les unités lexicales, il est évident qu'elles peuvent être polysémiques au même titre que les morphèmes.

Goldberg prend l'exemple de la construction ditransitive en anglais qui implique typiquement que l'argument agent cause le transfert d'un objet jusqu'à un bénéficiaire. Ce cas de transfert réussi (*actual successful transfert*) est le sens le plus fréquent de cette construction. Mais en réalité, plusieurs classes de verbes, associées à cette construction, font apparaître différentes variantes de ce sens basique. Fort de ces différences, la meilleure représentation de la sémantique requise est une classe de sens liés, c'est-à-dire que la forme ditransitive doit être associée à un ensemble de sens reliés de manière systématique. Cette construction peut être vue comme un cas de polysémie constructionnelle : la même forme est liée à des sens différents mais corrélés.

Goldberg ajoute qu'une analyse de la polysémie des constructions implique la reconnaissance du statut spécial d'un sens central ou basique de la construction. Ainsi Goldberg propose de donner à la construction ditransitive anglaise comme sens central le sens de 'transfert réussi', c'est-à-dire le sens à l'œuvre dans une scène où quelqu'un fait que quelqu'un reçoive quelque chose. De la même manière, à chaque construction on peut associer une scène humaine pertinente. Goldberg justifie cette notion de scène humaine pertinente centrale entourée de scènes dérivées en s'appuyant sur Langacker (1987), pour qui la langue est structurée autour de certains archétypes conceptuels.

Certains aspects différenciés de notre expérience, qui sont récurrents et saillants, apparaissent comme des archétypes que l'on utilise pour structurer notre appréhension du monde, autant que possible. A partir du moment où la langue est le moyen par lequel nous décrivons nos expériences, il est naturel que de tels archétypes soient repérés comme des valeurs prototypiques de constructions linguistiques basiques. Langacker suggère alors que ces archétypes sont déformés dans différentes directions par extension à partir des occurrences prototypiques parce que nous avons tendance à interpréter le nouveau ou moins familier en le référant à ce qui est déjà établi. C'est un mécanisme qui permet de limiter le nombre des unités conventionnelles.

Les rôles actanciels associés aux verbes

Le cadre sémantique d'un verbe inclut la détermination des 'rôles actanciels' (*participant roles*) que Goldberg distingue des rôles associés à la construction, appelés 'rôles argumentaux' (*argument roles*).

On peut, par des méthodes simples, découvrir quel est le nombre et le type des rôles actanciels impliqués par le cadre sémantique associé à un verbe donné. Prenons les deux exemples suivants :

(10) No kicking occurred. *(Il n'y a pas eu de coup de pied de donné)*

(11) No sneezing occurred. *(il n'y a pas eu d'éternuement)*

Dans l'énoncé (10), l'interprétation conduit à évoquer une scène (par ailleurs niée) qui comporte deux actants : le donneur et le receveur du coup de pied. On en déduit donc que le cadre sémantique de *kick* comporte deux actants, l'agent et le patient. En revanche, dans (11), la scène niée ne comporte qu'un actant, « l'éternueur », ce qui implique que le cadre sémantique de *sneeze* ne comporte qu'un actant.

Pour chaque verbe, on doit aussi déterminer lesquels de ces rôles doivent être obligatoirement présent (*profiled*), c'est-à-dire, dans la terminologie de Langacker, au premier plan sur la scène (par la suite nous avons traduit *profiled* par *inhérent* et *profiling* par *profil*). Les rôles actanciels *inhérents* au verbe doivent obligatoirement être

présents dans l'énoncé. Goldberg détaille sa description avec l'exemple des verbes *to rob* et *to steal* qui se traduisent tous les deux par le verbe *voler*. Ces deux verbes peuvent paraître synonymes, mais ils diffèrent dans leurs constructions syntaxiques, comme le montrent les différences d'acceptabilité des exemples suivants :

(12) Jesse robbed the rich *(*of all their money*).*

(13) *Jesse robbed a million dollars *(*from the rich*).*

(14) Jesse stole money *(*from the rich*).*

(15) *Jesse stole the rich *(*of money*).*

Goldberg propose de distinguer *rob* et *steal* en leur attribuant les mêmes rôles actanciels, mais ayant un *profil* différent (les rôles actanciels *inhérents* au verbe sont mis en **gras**) :

rob: < **thief target** goods >

steal: < **thief** target **goods** >

Pour *rob*, le voleur et le volé sont au premier plan, alors que ce sont le voleur et ce qui est volé dans le cas de *steal*.

Les rôles argumentaux associés aux constructions

Les constructions, au même titre que les unités lexicales, spécifient aussi des rôles *inhérents* : tout rôle argumental lié à une relation grammaticale directe, sujet ou objet direct (c'est-à-dire sans préposition), doit être « *profilé* ». Ainsi, la construction ditransitive comporte trois rôles argumentaux *inhérents*. Elle est associée au schéma 'X (agent) fait que Y (bénéficiaire) reçoive Z (patient)', que Goldberg propose de représenter par :

CAUSE-RECEIVE<**agt rec pat**>

On doit pouvoir spécifier pour chaque construction de quelle manière les verbes doivent être combinés avec elle. On doit aussi pouvoir spécifier la classe de verbes

qu'elle peut accepter, et comment le type d'événement désigné par le verbe est intégré dans le type d'événement désigné par la construction.

La fusion des rôles actanciels et des rôles argumentaux

La fusion (*blending*) est un processus analogue au mécanisme d'unification, tel qu'il est utilisé dans différents formalismes syntaxiques (qu'on appelle justement grammaires d'unification). Il consiste à prendre en compte simultanément les contraintes sémantiques des rôles actanciels associés au verbe, et les contraintes sémantiques des rôles argumentaux de la construction. Choisir quel rôle actanciel est fusionné avec quel rôle argumental est déterminé par deux principes :

- le principe de cohérence sémantique : seuls les rôles pouvant être sémantiquement compatibles peuvent être fusionnés. Deux rôles r1 et r2 sont sémantiquement compatibles si r1 peut être interprété comme une instance de r2, ou si r2 peut être interprété comme une instance de r1.

- le principe de correspondance : Tout rôle actanciel lexicalement *inhérent* au verbe doit être fusionné avec un rôle argumental *inhérent* à la construction.

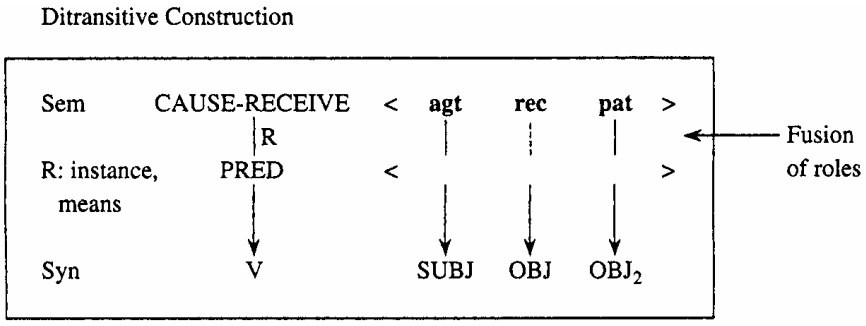


Figure 7.1 : Représentation de la construction ditransitive

La figure 7.1 montre un couplage entre un niveau sémantique et un niveau syntaxique de fonctions grammaticales. Dans le cas typique, les rôles actanciels associés au verbe peuvent être associés un par un aux rôles argumentaux associés à la

construction. C'est le cas de *hand* pour la construction ditransitive (tendre quelque chose à quelqu'un. Cf. figure 7.2).

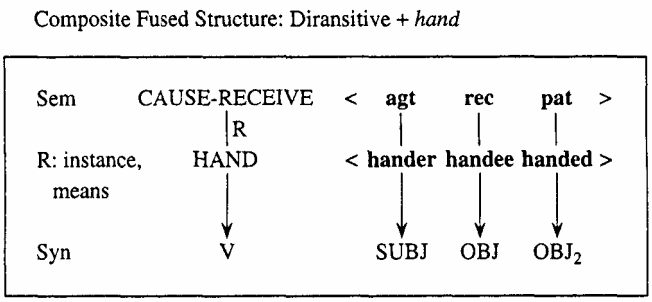


***Figure 7.2 : Représentation de la construction ditransitive couplée avec le verbe* hand**

Non concordance des rôles

C'est bien sûr dans les cas où il n'y a pas concordance des rôles actanciels et argumentaux que la théorie prend tout son sens. Par exemple, le verbe *bake* ne comporte dans son cadre sémantique que deux rôles actanciels, l'agent et le patient. Quand il est utilisé dans une construction ditransitive, comme dans (6) *She baked him a cake*, on se trouve donc dans un cas de non concordance puisque le bénéficiaire n'est pas présent dans la sémantique du verbe.

Pour illustrer le mécanisme de fusion dans ces cas, nous allons présenter plus en détails l'une des constructions que Goldberg a particulièrement étudiées : la construction de mouvement induit.

7.2.2 La construction de mouvement induit (caused motion construction)

Goldberg a utilisé cette construction afin de pouvoir expliquer certains cas dans lesquels l'interprétation sémantique ne peut être attribuée ni au verbe principal, ni au résultat de la compositionnalité sémantique. Elle peut être définie structurellement comme suit :

[SUBJ [V OBJ OBL]]

où V est un verbe non statique et OBL est un syntagme directionnel.

Cette définition permet de couvrir les types d'expressions suivantes :

(16) Franck sneezed the tissue off the table *(Frank a fait valser le mouchoir de la table en éternuant)*

(17) Mary urged Bill into the house *(Marie a insisté pour que Bill entre dans la maison)*

(18) Sue let the water out of the bath *(Sue a laissé l'eau s'écouler de la baignoire)*

(19) Sam helped him into the car *(Sam l'a aidé à entrer dans la voiture)*

(20) They sprayed the paint onto the wall *(Ils ont pulvérisé la peinture sur le mur)*

L'argument 'agent' est la cause directe du mouvement de l'argument 'patient' (appelé aussi ici 'thème'), et le mouvement est spécifié par le syntagme directionnel, ce que Goldberg représente par la formule : X CAUSES Y to MOVE Z.

L'existence de la construction

Pour prouver que la construction de mouvement induit est bien une construction distincte à part entière, Goldberg doit montrer que sa sémantique n'est pas issue de la dérivation compositionnelle d'autres constructions existantes dans la grammaire. Par exemple, il est nécessaire de montrer qu'elle n'est pas dérivée de la simple union des unités lexicales qui la composent.

Plusieurs observations dans la littérature conduisent à établir que le verbe isolé ne peut encoder la sémantique de cette construction. Fillmore (1971) a montré que de nombreux verbes ne sont pas causatifs (l'action n'est pas forcément intentionnelle) lorsqu'ils ne sont pas dans cette construction. C'est le cas par exemple, pour *kick* et *hit* dans les énoncés suivants :

(21) Joe kicked the wall

(22) Joe hit the table

Pourtant quand ces verbes sont employés dans une construction de mouvement induit, une interprétation causative, intentionnelle, apparaît :

(23) Joe kicked the dog into the bathroom. *(Joe a fait entrer le chien à coups de pied dans la salle de bain)*

(24) Joe hit the ball across the field. *(Joe, en tapant dans la balle, lui a fait traverser le terrain)*

C'est aussi le cas pour les verbes qui ne comportent pas nécessairement de notion de mouvement indépendamment de cette construction :

(25) Franck squeezed the ball *(Franck a pressé la balle)*

(26) Franck squeezed the ball through the crack *(Franck a fait entrer la balle dans la fissure en la pressant)*

La balle ne bouge pas nécessairement dans (25), contrairement à (26).

Ces illustrations tendent à démontrer qu'apparaît pour chacun de ces verbes un nouveau sens de mouvement intentionnel. La construction de mouvement induit peut être représentée comme suit :

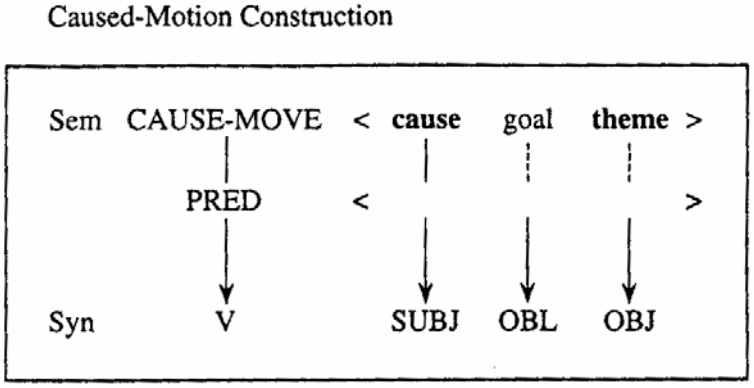


Figure 7.3 : Représentation de la construction de mouvement induit

Les variantes de cette construction

Goldberg décrit différentes variantes de cette construction : la variante 'd'autorisation' avec les verbes *allow*, *let*, *free*, *release*, ou à l'inverse, la variante 'd'interdiction' avec *lock*, *keep*, ou encore la variante 'd'acte communicatif' dont voici des illustrations :

(27) Sam ordered him out of the house

(28) Sam asked him into the room

(29) Sam sent him to the market

Le mouvement n'est pas strictement impliqué dans ces variantes. L'énoncé (27) n'implique pas forcément que la personne est sortie de la maison. On peut simplement dire qu'au niveau pragmatique, le mouvement est impliqué par « la condition de satisfaction » (Searle, 1983) associée à l'action évoquée par le prédicat. comme on l'a dit, pour Goldberg, le cas présenté initialement, qui implique une causalité active et un mouvement direct, est le sens de base de cette construction, les autres sens étant considérés comme des extensions de sens correspondant à une certaine polysémie de la construction. (cf. annexe C pour un schéma descriptif des différentes variantes de cette construction).

7.3. Discussions de Goldberg et des grammaires constructionnelles

Nous n'avons bien sûr pas présenté l'ensemble du travail de Goldberg dans ce chapitre, mais nous avons plutôt essayé d'en donner l'esprit général, notamment sa manière d'appréhender les interactions entre verbe et construction. Nous proposons pour finir de discuter certains points précis développés par Goldberg, en particulier sur le formalisme qu'elle propose. Nous verrons aussi quelques critiques qui touchent spécifiquement la construction de mouvement induit.

7.3.1 Sens de base des constructions

Tout d'abord, en ce qui concerne la représentation de la polysémie des constructions, Goldberg semble dire que le fait qu'une construction soit polysémique implique nécessairement l'existence d'un sens de base de cette construction. Nous avons montré dans le chapitre 2 qu'il était possible de rendre compte de la polysémie d'unités lexicales en reliant les sens par des ressemblances de famille à la Wittgenstein. Appliquer cette structure aux constructions permettrait d'éviter le choix toujours discutable d'un sens « premier ». Pour la construction ditransitive, ce sens de base correspond pour Goldberg au sens de 'transfert direct accompli', alors qu'apparaissent parmi les sens dérivés 'condition de satisfaction pour un transfert direct', 'intention de transfert direct', 'permission de transfert direct', 'transfert communicatif', etc. Sur quels

critères Goldberg a-t-elle choisi le 'transfert direct accompli' comme noyau de sens ? Qu'il s'agisse de critères de diachronie (premiers sens attestés historiquement), de fréquence d'usage, ou encore d'acquisition par l'enfant, cela mériterait d'être appuyé sur des études précises, dont on peut douter qu'elles aient effectivement été menées. En l'absence de données avérées, un tel choix, de par son caractère subjectif, reste très discutable, « l'intuition » des locuteurs étant souvent prise en défaut dans ce type de phénomène. Cette proposition d'un sens de base est d'ailleurs d'autant plus critiquable qu'elle n'est pas indispensable au modèle : il n'y a aucune raison de privilégier un sens ou un autre *a priori* dans le calcul des interactions sémantiques entre verbe et construction.

7.3.2 Correspondance de rôles

Une deuxième remarque concerne la combinaison des rôles actanciels et des rôles argumentaux. Goldberg présente le cas où il y a concordance complète des rôles (notamment avec le couple formé par la construction ditransitive et le verbe *hand*), mais aussi d'autres cas où la concordance n'est pas parfaite mais où elle peut être adaptée en prenant en compte une certaine polysémie du verbe concerné. Elle prend comme exemple le verbe *lease* dans les énoncés suivants :

(30) Cecile leased the apartment from Ernest *(Cécile est la locataire)*

(31) Ernest leased the apartment to Cecile *(Ernest est le propriétaire)*

Combien de rôles actanciels sont *inhérents* pour le verbe *lease* ? Il n'est pas possible de dire que *lease* n'a qu'un rôle activé, la propriété, car le verbe ne peut pas être utilisé avec un seul actant. L'énoncé suivant est en effet inacceptable :

(32) *The appartment leased

Pour représenter ces cas, Goldberg propose alors de considérer que le verbe a deux sens distincts, qui correspondent à deux profils différents :

lease$_1$: <**tenant property** landlord>

lease$_2$: <tenant **property landlord**>

Goldberg insiste sur le fait que *lease* n'a qu'un seul cadre sémantique <tenant property landlord> , et que seuls les rôles actanciels inhérents changent. Mais le problème, c'est qu'il n'y a pas seulement une question de profil de rôle actanciel, mais aussi de différence de rôle argumental dans la phrase. En (30) le locataire est sujet de la phrase alors qu'en (31) il est objet indirect, et inversement pour le propriétaire. Pour traiter ces cas, il faut donc d'abord prendre en compte les différences de construction, à savoir la différence de préposition du complément indirect (*from* et *to*) : c'est ce qui va permettre d'établir la bonne correspondance entre rôles actanciels et rôles argumentaux. Autrement dit, il faut d'abord associer à son rôle argumental le rôle actanciel non inhérent, ce qui relativise l'intérêt de cette notion de profil actanciel (elle ne joue de fait aucun rôle dans le processus de fusion).

Cette remarque concerne une partie minime de la description de Goldberg néanmoins elle soulève les limites de ce formalisme, notamment concernant les interactions entre constructions polysémiques et verbes polysémiques. On trouvera dans Michaelis (2003) une proposition plus élaborée de classification des constructions suivant le type d'opérations impliquées lors de la fusion du sens du verbe et du sens de la construction, qui prend notamment en compte ces phénomènes de polysémie.

7.3.3 La construction de mouvement induit

Cette construction est celle qui a le plus marqué les travaux de Goldberg, et c'est aussi celle qui est le plus sujette à discussion. Comme nous l'avons dit, pour prouver l'existence de cette construction, Goldberg doit montrer que sa sémantique n'est pas issue de la dérivation compositionnelle, d'autres constructions existant dans la grammaire. En particulier il est nécessaire de montrer que ce ne sont pas les unités elles-mêmes qui portent les effets de sens que l'on impute à la construction. Goldberg cherche à le démontrer en ce qui concerne la notion de mouvement en contrastant des exemples tels que (25) et (26) :

(25) *Franck squeezed the ball* (Franck a pressé la balle)

(26) *Franck squeezed the ball through the crack* (Franck a fait entrer la balle dans la fissure)

Or une autre explication de cette différence entre ces deux énoncés pourrait être la présence de la préposition *through* (à travers). Cette préposition est en elle-même porteuse de la notion de mouvement. Ainsi, la préposition peut être déterminante pour l'interprétation globale de la phrase. Considérons l'énoncé suivant :

(33) Franck squeezed the ball on the crack *(Franck a coincé la balle dans la fissure en la pressant)*

Il n'y a pas d'évocation de mouvement, bien qu'il s'agisse de la même construction. La plupart des exemples de mouvement induit font intervenir, de la même façon, une préposition qui implique déjà à elle seule un mouvement : *into*, *onto*, *off*, que l'on peut contraster avec leurs équivalents statiques *in*, *on*, *out of*. On pourrait alors considérer que l'évocation de mouvement n'est due qu'à la combinaison des unités qui composent la phrase, si l'on inclut les unités grammaticales. Si tel est le cas, la construction de mouvement induit ne serait plus considérée comme une construction au sens de Goldberg.

Dans le même ordre d'idées, Paul Kay (2001) apporte aussi quelques critiques concernant cette construction. Il explique notamment que la plupart des exemples donnés par Goldberg pour appuyer l'existence de cette construction ne correspondent pas à une construction mais à des additions successives d'arguments, en s'appuyant sur les exemples suivants :

(34) The top was spinning *(le bouchon tournoyait)*

(35) Kim was spinning the top *(Kim faisait tourner le bouchon)*

(36) The top was spinning off the table *(le bouchon quittait la table en tournoyant)*

(37) Kim was spinning the top off the table *(Kim faisait quitter la table au bouchon en le faisant tourner)*

De (34) à (35) on a l'ajout d'un agent causatif ; de (34) à (36) l'ajout d'un argument de direction au verbe intransitif ; dans ce cas, (37) peut être considéré comme

la combinaison de (35) et (36). Kay admet cependant que ces additions successives d'arguments ne sont pas possibles pour tous les exemples de construction de mouvement induit.

7.3.4 Conclusion

L'approche constructionnelle de la grammaire est celle qui correspond le mieux à notre conception du calcul du sens d'un énoncé, pour plusieurs raisons. D'abord le refus d'une séparation brutale entre lexique et syntaxe nous semble tout à fait justifié par l'existence de nombreux phénomènes pour lesquels il est difficile de trancher entre un apport de la construction syntaxique et des unités linguistiques impliquées. C'est le cas, comme on vient de le voir (ou d'en discuter), de la construction de mouvement induit. De plus, la grammaire constructionnelle confère un rôle nouveau aux constructions syntaxiques en leur attribuant un sens propre, ce qui permet effectivement de traiter dans un même cadre théorique les apports des différentes unités (lexicales ou syntaxiques).

Enfin ce cadre permet de considérer les constructions syntaxiques comme des unités polysémiques, ce qui renforce l'homogénéité d'un modèle comme le nôtre qui accorde une place essentielle à la polysémie.

Les travaux de Goldberg nous intéressent plus particulièrement parce qu'elle travaille sur les structures argumentales, qu'elle approfondit la description des relations entre verbes et constructions, et enfin parce qu'elle propose un formalisme permettant de rendre compte des différents cas d'interaction entre verbes et constructions. Nous ne retiendrons pas nécessairement le formalisme lui-même, qui, nous l'avons vu, peut être discutable par certains aspects, mais nous retiendrons ce qu'il a permis de mettre en évidence, c'est-à-dire les différents cas d'interaction entre verbe et construction.

8. Les constructions verbales : calcul de l'influence du co-texte syntaxique

Les grammaires constructionnelles couvrent un large domaine de l'interprétation linguistique. Nous proposons d'utiliser cette grammaire comme fondement théorique pour notre approche des constructions syntaxiques. Reprenons les caractéristiques que nous allons exploiter :

1. les constructions sont porteuses d'un sens intrinsèque
2. les constructions sont polysémiques
3. le fait de proposer une représentation commune pour le lexique et les constructions syntaxiques, et refuser une séparation brutale entre ces deux types d'unité linguistique.

Nous utilisons l'hypothèse (1) pour faire le postulat de travail suivant : le sens d'un verbe dans une construction donnée est proche du sens des synonymes de ce verbe que l'on retrouve dans la même construction. Cela revient à dire que *compter* avec la construction *V sur SN* est synonyme de *tabler* et *s'appuyer* parce qu'on peut dire *tabler sur SN*, *s'appuyer sur SN,* mais pas synonyme de *contenir* ou *énumérer* puisque les énoncés **contenir sur SN* ou **énumérer sur SN* ne sont pas corrects. Il est aisé de trouver des contre-exemples à ce postulat. Le sens de *compter* dans c*ompter sur quelque chose* est proche de *espérer*, mais la construction n'est pas la même : on dit *espérer quelque chose* et non **espérer sur quelque chose*. Ce postulat n'est donc pas un principe général mais plutôt une tendance qui se vérifie dans les faits. Nous allons voir qu'il est possible d'exploiter ce postulat en termes statistiques.

L'idée est alors de calculer, au même titre que pour le co-texte lexical, le degré d'affinité d'une construction syntaxique avec le verbe étudié. L'intérêt de ce choix étant d'une part de respecter le modèle théorique de la construction dynamique du sens et l'hypothèse (3) des grammaires constructionnelles, puisque cela consiste à considérer l'influence du co-texte syntaxique de la même manière que les autres unités co-textuelles. D'autre part, les fonctions potentielles obtenues pourront rendre compte de sens précis, ambigus ou indéterminé, donc respecter l'hypothèse (2) considérant les constructions comme polysémiques.

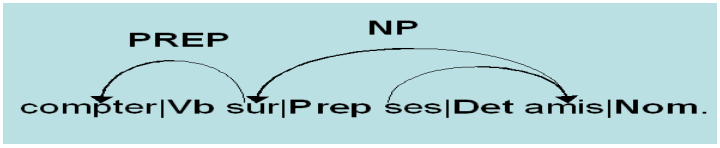
Nous commencerons par présenter notre méthode de calcul. Nous en profiterons pour décrire plus en détail la construction progressive de la fonction potentielle, et notamment les différentes étapes qui permettent de rendre notre calcul robuste. Nous décrirons ensuite la partie logicielle que nous avons développée. Les deux parties suivantes correspondent à deux évaluations de notre modèle. Nous avons déjà évoqué la difficulté d'évaluer un modèle de désambiguïsation (chapitre 3, § 3.2). Cette difficulté est d'autant plus grande avec une représentation continue du sens puisque nous nous refusons à baser notre évaluation sur un corpus annoté sémantiquement. Nous proposons alors deux types d'évaluations que nous considérons comme complémentaires. Nous terminerons cette partie sur un modèle de combinaison entre co-texte lexical et co-texte syntaxique.

8.1. Méthode

Pour cette tâche nous utiliserons le corpus que nous avons nommé LM3. Rappelons que ce corpus est constitué de tous les articles du journal *Le Monde* sur trois ans (1994-1996), soit 11 millions de mots[44]. L'analyseur syntaxique Syntex (Bourigault et Fabre, 2000) est utilisé pour extraire de ce corpus des ensembles de mots ou

[44] Nous remercions Didier Bourigault de nous avoir autorisé à travailler à partir de son corpus.

syntagmes[45], structurés par des relations de dépendance syntaxique. Par exemple l'énoncé *« compter sur ses amis »,* sera analysé par Syntex de la manière suivante (cf. annexe D pour une illustration du type de sortie proposée par Syntex):



Il est alors possible d'extraire de ce corpus le nombre d'occurrence d'un verbe dans une construction donnée, même complexe (*X demande Y à Z*). Cela nous permet d'appliquer la même méthode que pour le co-texte lexical. Pour chaque synonyme du verbe étudié (obtenu par le D.E.S.), on cherche à savoir s'il est employé dans la construction étudiée. Pour cela on calcule la fréquence relative de chaque synonyme employé dans la construction étudiée, et on compare cette fréquence à la fréquence théorique à l'aide du degré d'affinité que nous avons présenté dans le paragraphe 5.2.2.

Ainsi, de la même manière que pour le co-texte lexical, on aura a(k,j) le degré d'affinité de la construction s_j avec la clique c_k, et la fonction potentielle associée à la construction s_j, avec les coordonnées (x,y) sera :

$$f_j(x,y) = \max\left(0, \sum_{k=1}^{c} a(k,j)\, e^{-\frac{(x_k - x)^2 + (y_k - y)^2}{\delta^2}}\right)$$

(avec les mêmes notations que dans le chapitre 5 § 5.2.4)

Nous proposons d'illustrer ce processus en décrivant quelques étapes du calcul ainsi que les résultats obtenus pour la construction prépositionnelle *V + SP (sur SN)* pour le verbe *compter*. Puis les résultats obtenus pour la construction transitive. Ci-

[45] Syntex permet de considérer certains syntagmes tels que *chef d'état*, *groupe financier* ou *Parc des Princes* comme des unités à part entière.

dessous un récapitulatif des fréquences exploitées pour le calcul ainsi qu'un rappel de la structure de l'espace sémantique de *compter* (figure 8.1), que nous avons déjà présenté dans le chapitre 4.

	Fréquences
Compter	5 491
V SP (sur SN)	43 043
V OBJ (SN)	638 363
Compter SP (sur SN)	664
Compter OBJ (SN)	2 058

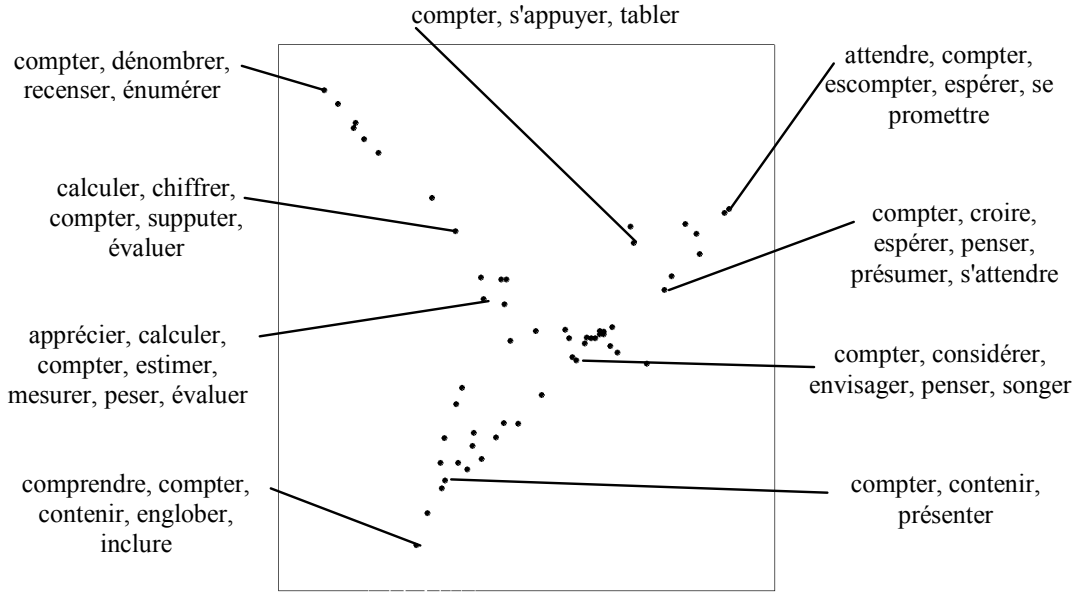


***Figure 8.1 : Espace sémantique du verbe* compter**

peser	955
s'appuyer	771
fonder	768
compter	664
prendre	227
tabler	178
exister	93
présenter	76
attendre	64
entendre	64
projeter	60
comprendre	48
calculer	40
payer	32
introduire	31
croire	24
marquer	20
regarder	19
mesurer	11
espérer	11
estimer	11
chiffrer	10
penser	9
recenser	6
facturer	5
escompter	4
dater	4
posséder	4
importer	4
examiner	4
envisager	4
épargner	3
présumer	3
apprécier	3
considérer	3
supputer	2
dénombrer	1
réputer	1
contenir	1
évaluer	1

***Figure 8.2 : Liste de synonymes de* compter *par ordre de fréquence d'emploi avec la construction* V sur SN**

s'appuyer	0,97385
fonder	0,84754
projeter	0,60618
facturer	0,40421
chiffrer	0,34872
tabler	0,32541
exister	0,31977
calculer	0,18736
payer	0,16794
introduire	0,12723
peser	0,12681
supputer	0,1117
recenser	0,092615
épargner	0,081549
attendre	0,076122
entendre	0,074746
présenter	0,058733
mesurer	0,051762
escompter	0,050496
prendre	0,044547
dater	0,0365
comprendre	0,035021
marquer	0,031424
regarder	0,021613
présumer	0,015351
dénombrer	0,015134
posséder	0,013809
importer	0,012541
examiner	0,0094034
compter	0,0093652
croire	0,0083343
apprécier	0,0082551
espérer	0,0075734
réputer	0,0062801
envisager	0,0060407
contenir	0,0032356
penser	0,0025581
évaluer	0,0020269
estimer	0,0012285
considérer	0,0011135

***Figure 8.3 : Liste de synonymes de* compter *par ordre de degré d'affinité avec la construction* V sur SN**

compter ; s'appuyer ; tabler	1,3086
compter ; fonder ; tabler	1,1823
calculer ; chiffrer ; compter ; supputer ; évaluer	0,65917
compter ; envisager ; penser ; projeter ; se proposer ; songer	0,62415
calculer ; compter ; estimer ; mesurer ; peser ; supputer ; évaluer	0,49024
compter ; facturer	0,41357
compter ; escompter ; espérer ; tabler	0,39285
apprécier ; calculer ; compter ; estimer ; mesurer ; peser ; évaluer	0,3868
compter ; exister	0,32913
compter ; estimer ; examiner ; peser ; supputer ; évaluer	0,26053
compter ; introduire ; présenter	0,19533
compter ; escompter ; espérer ; supputer	0,17913
compter ; payer	0,1773
compter ; considérer ; estimer ; examiner ; penser ; peser ; regarder	0,17209
apprécier ; compter ; estimer ; examiner ; peser ; évaluer	0,15709
apprécier ; compter ; considérer ; estimer ; examiner ; peser	0,15617
compter ; importer ; introduire	0,14914
compter ; entrer en ligne de compte ; importer ; peser	0,14871
attendre ; compter ; escompter ; espérer ; se promettre	0,14356
compter ; inclure ; introduire	0,1366

***Figure 8.4 : Liste des cliques contenant* compter *par ordre de degré d'affinité avec la construction* V sur SN**

Les figures 8.2, 8.3 et 8.4 montrent quelques étapes du processus. La figure 8.2 présente la liste des synonymes de *compter* dans l'ordre décroissant de leur fréquence d'emploi avec la construction *V sur SN*. La figure 8.3 présente la même liste de verbes mais cette fois dans l'ordre décroissant de leur degré d'affinité. On peut noter que l'influence de verbes tels que *peser* ou *prendre* qui sont très fréquents dans le corpus, est amoindrie. A l'inverse *facturer, chiffrer* et *tabler* sont mis en valeur. La figure 8.4 présente la liste des cliques dans l'ordre décroissant de leur degré d'affinité. Le fait de considérer le degré d'affinité des cliques et non des synonymes pour le calcul de la fonction potentielle rend l'approche bien plus robuste. En effet une clique contenant beaucoup de synonymes ayant un degré d'affinité élevé sera privilégiée à une autre n'ayant que quelques synonymes à degré d'affinité élevé. C'est ainsi que « *compter, s'appuyer, tabler* » et « *compter, fonder, tabler* » sont les premières cliques alors que *tabler* n'est qu'en 6ème position.

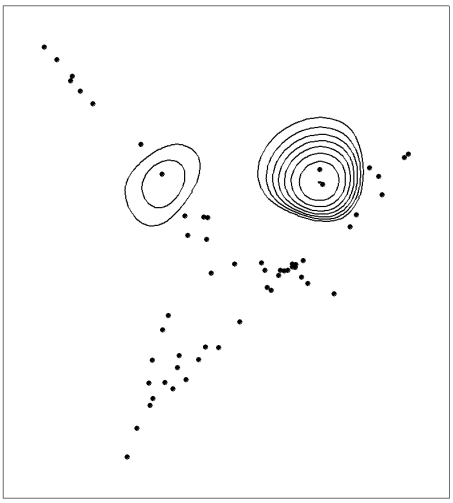


***Figure 8.5 : Fonction potentielle de la construction* V sur SN *sur l'espace sémantique de* compter**

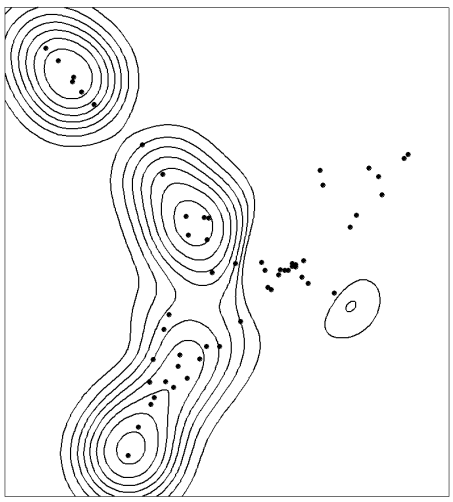


***Figure 8.6 : Fonction potentielle de la construction* V SN *sur l'espace sémantique de* compter**

Les figures 8.5 et 8.6 représentent les fonctions potentielles des constructions respectivement *V sur SN et V SN* sur le verbe *compter*. On peut observer que ces constructions activent des régions très différentes de l'espace sémantique de *compter*. La

construction *V sur SN* sélectionne presque exclusivement le sens 'tabler' (*compter sur quelqu'un*), avec une petite incursion du côté du dénombrement (à l'œuvre dans l'emploi *compter sur ses doigts*). Cette représentation de la construction *compter sur SN* met en évidence le fait que l'analyseur syntaxique ne distingue pas les compléments transitifs indirects dits essentiels (compter sur quelqu'un), des compléments intransitifs dits circonstanciels (compter sur ses doigts). Au passage, cela donne aussi une explication au fait que les dictionnaires ne s'accordent pas sur la manière de classer la construction *compter sur SN* (cf. § 4.1.2).

On peut observer que la clique « *compter ; envisager ; penser ; projeter ; se proposer ; songer* », qui a pourtant un degré d'affinité de 0,62415 (4ème place dans la figure 8.4) n'a activé aucune zone dans l'espace sémantique. Cela illustre le fait que notre fonction potentielle ne met en avant que les zones ayant beaucoup de cliques à degré d'affinité élevé et non les zones n'ayant qu'une clique à degré d'affinité fortement élevé.

En revanche, la construction transitive *V SN* (figure 8.6) n'est pas très sélective. On trouve en effet plusieurs sens de *compter* dans cette construction : *compter les moutons*, *ce village compte 2 000 âmes*, *sans compter les enfants*, *compter ses sous*, etc. Elle exclut tout de même une bonne partie de l'espace sémantique (notamment les sens 'importer', 'considérer', 'espérer' : *il compte beaucoup pour moi*, *je compte aller le voir*, etc.).

8.2. Développement logiciel

En travaillant sur un nouveau corpus, nous nous sommes retrouvé face à de nouveaux problèmes logiciels. En effet, Frantext n'était peut-être pas un corpus analysé syntaxiquement, mais les fonctions de calcul de fréquences étaient déjà en place à l'ATILF. Pour le corpus LM3, nous avons dû construire ces fonctions. Une autre difficulté était la complexité nouvelle des requêtes. Nous sommes passé de requêtes du type « Quelle est la fonction potentielle du co-texte X ?» à « Quelle est la fonction

potentielle du syntagme prépositionnel rattaché au verbe, qui est introduit par la préposition P et dont la tête nominale de ce syntagme est un nom quelconque ?». Les étapes de ce logiciel sont :

- Entrée de la requête de l'utilisateur à l'aide d'une interface
- Construction automatique d'un programme Perl correspondant à cette requête
- Exécution du programme Perl et rapatriement des fréquences calculées
- Calcul et affichage de la fonction potentielle correspondante

Nous présenterons ici les deux premières parties. Les deux dernières étant analogues au logiciel décrit dans le paragraphe 6.1.

L'interface a nécessité beaucoup d'attention puisqu'il ne fallait plus traiter la requête en linéaire, c'est-à-dire les mots les uns à la suite des autres, mais en terme de dépendance. Toute la difficulté est d'avoir une interface compréhensible par l'utilisateur, tout en étant la plus complète possible.

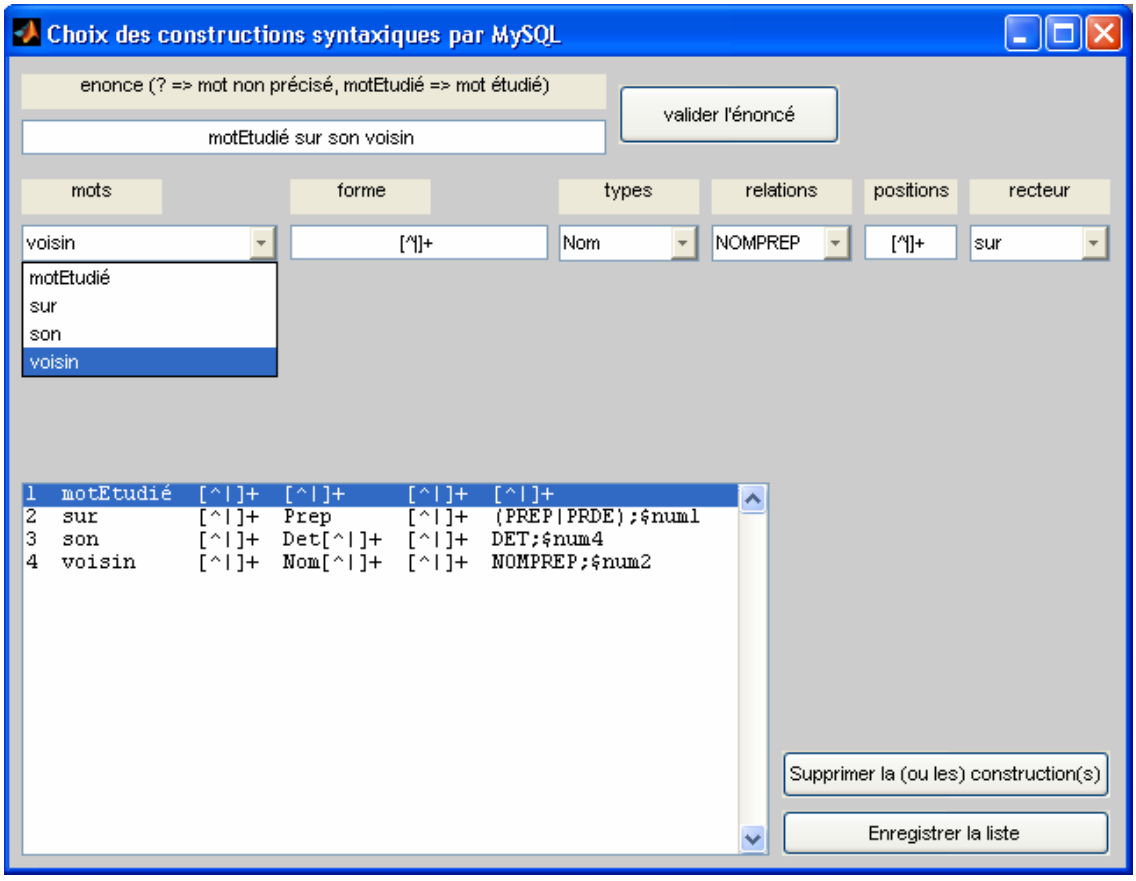


Figure 8.7 : Interface pour les requêtes Syntex

Le principe de l'interface présentée en figure 8.7 est le suivant : l'utilisateur entre la construction qu'il souhaite étudier. Ici, l'utilisateur travaille sur le verbe *compter* (qui a été précisé avant) et s'intéresse à la construction *compter sur son voisin* (notons que *compter* est remplacé par « motEtudié », ceci afin de faciliter l'interprétation de l'énoncé par le programme). Si l'utilisateur avait voulu étudier la construction *compter sur SN*, il aurait entré « motEtudié sur ? », où le point d'interrogation signifie que l'on ne souhaite pas préciser la lexie.

La deuxième étape consiste à préciser, pour chaque élément de la construction demandée, les caractéristiques de cet élément et les relations qu'il entretient avec les autres éléments. Concernant ses caractéristiques, il est possible de préciser sa forme, son type (catégorie grammaticale), sa position dans l'énoncé par rapport au mot vedette. On peut ensuite déterminer si ce mot est régi par une relation syntaxique, de quel type est cette relation et avec quel mot. Chaque nouvelle précision proposée par l'utilisateur est ajoutée dans la liste récapitulative en bas à gauche de la fenêtre. Cette liste est assez difficile à déchiffrer puisqu'elle mélange des expressions Perl avec des entités de Syntex, mais elle permet pour les initiés de vérifier que leur requête a bien été enregistrée. Il est important de préciser que chacune des caractéristiques que nous venons de présenter ne doit pas obligatoirement être donnée par l'utilisateur. Autrement dit, l'utilisateur entre ce qu'il connaît ou ce qu'il attend, et le programme s'en accommode.

Un fois que la requête est enregistrée par l'utilisateur, le logiciel doit la traduire en requête Perl. Afin de simplifier l'implémentation, nous avons développé un programme récursif permettant d'écrire la requête Perl à la manière métaphorique d'un emboîtement de poupées russes. La figure 8.8 représente l'algorithme de ce programme récursif, et nous proposons en annexe B.2 la requête Perl construite pour le calcul de fréquence de la construction *compter sur son voisin*.

```
fonction prog = constrProgramme(listMots,prog)
    %listMots: listes des mots de la requête avec leurs caractéristiques
    %prog : variable texte de stockage du programme Perl
    %En entrée, prog est vide

    Choix du mot m à traiter : Recherche d'un mot de listMots dont tous les
    recteurs ont déjà été traités
    listeMots = listeMots - m
    Prog = prog + début du programme concernant le mot m
    Si listeMots non égal zéro
        constrProgramme (listeMots,prog)
    Sinon
        Prog = prog + programme de test de validité de la construction
    Finsi
    Prog = prog + fin du programme concernant le mot m
```

Figure 8.8 : Fonction récursive pour la construction de requête Perl

Nous terminerons cette partie *logiciel* sur un développement plus récent qui n'a pas eu le temps d'aboutir à un système fini et qui consiste à ne plus travailler directement sur corpus mais à partir de bases SQL. En effet, la recherche directe sur corpus est très coûteuse en temps, et ce pour des requêtes en général assez redondantes. L'idée est alors de stocker sous format SQL le corpus afin de rendre plus efficace les requêtes. L'avantage de ce format est d'optimiser le stockage de l'information et d'améliorer les temps de calcul, du moins pour les requêtes les plus courantes.

8.3. Evaluation qualitative sur le verbe *jouer*

Cette première évaluation consiste à confronter la fonction potentielle calculée par notre modèle avec une étude linguistique détaillée. C'est-à-dire une étude proposant pour un verbe donné, de décrire tous les sens possibles en fonction de la construction dans laquelle il est employé. Pour un verbe *v* employé dans la construction *c*, le principe de l'évaluation sera de reporter les sens décrits par l'étude linguistique sur l'espace sémantique de *v*, puis d'observer le degré de superposition entre ces points de l'espace et la fonction potentielle de *c* que nous avons calculée. Ce mode d'évaluation est nécessairement restreint puisque cela nécessite une analyse sémantique approfondie

pour un verbe donné, et que cette analyse soit basée sur différentes constructions du verbe. L'étude sur les constructions prépositionnelles du verbe *jouer* faite par Pierre Cadiot (1999) correspond à ces critères. Cadiot y décrit les sens du verbe *jouer* dans plusieurs constructions par une série d'énoncé-types ainsi que quelques synonymes permettant de préciser le sens qu'il veut exprimer avec chacun de ces énoncés. Nous allons utiliser cette étude pour évaluer notre modèle sur les constructions *jouer sur SN, jouer avec SN, jouer à SN, jouer de SN* qui ont été décrites pas Cadiot. Pour chaque construction prépositionnelle évaluée, nous présentons la région activée par notre modèle pour cette construction sur l'espace sémantique de *jouer*, et nous reportons les sens proposés par Cadiot pour cette construction sur ce même espace. Nous commentons ensuite la superposition.

8.3.1 La construction prépositionnelle V sur SN

Le tableau ci-dessous décrit les différents sens de *jouer* proposés par Cadiot pour la construction *V sur SN*. La première colonne correspond aux différents énoncé-types proposés par Cadiot, la seconde correspond à une explication du sens donné par Cadiot pour le verbe *jouer*, enfin la dernière colonne correspond à des synonymes de *jouer* présents dans le dictionnaire des synonymes que nous utilisons et qui selon nous illustrent le sens que Cadiot donne à *jouer* dans chaque énoncé.

Enoncés illustratifs (Cadiot)	Précision du sens (Cadiot)	Synonymes correspondants
jouer sur un cheval	*mettre sa mise*	*miser, parier*
jouer sur les grains	*spéculer*	*spéculer, miser, parier*
la barque joue sur son ancre	*se mouvoir*	*se mouvoir, remuer*
jouer sur les nerfs de quelqu'un	*tirer profit de la nervosité*	*agiter, manier*
jouer sur plusieurs registres	*mettre en œuvre*	*agir, influer*
jouer sur les mots	*tirer parti des équivoques*	*feinter, rouler, tromper*

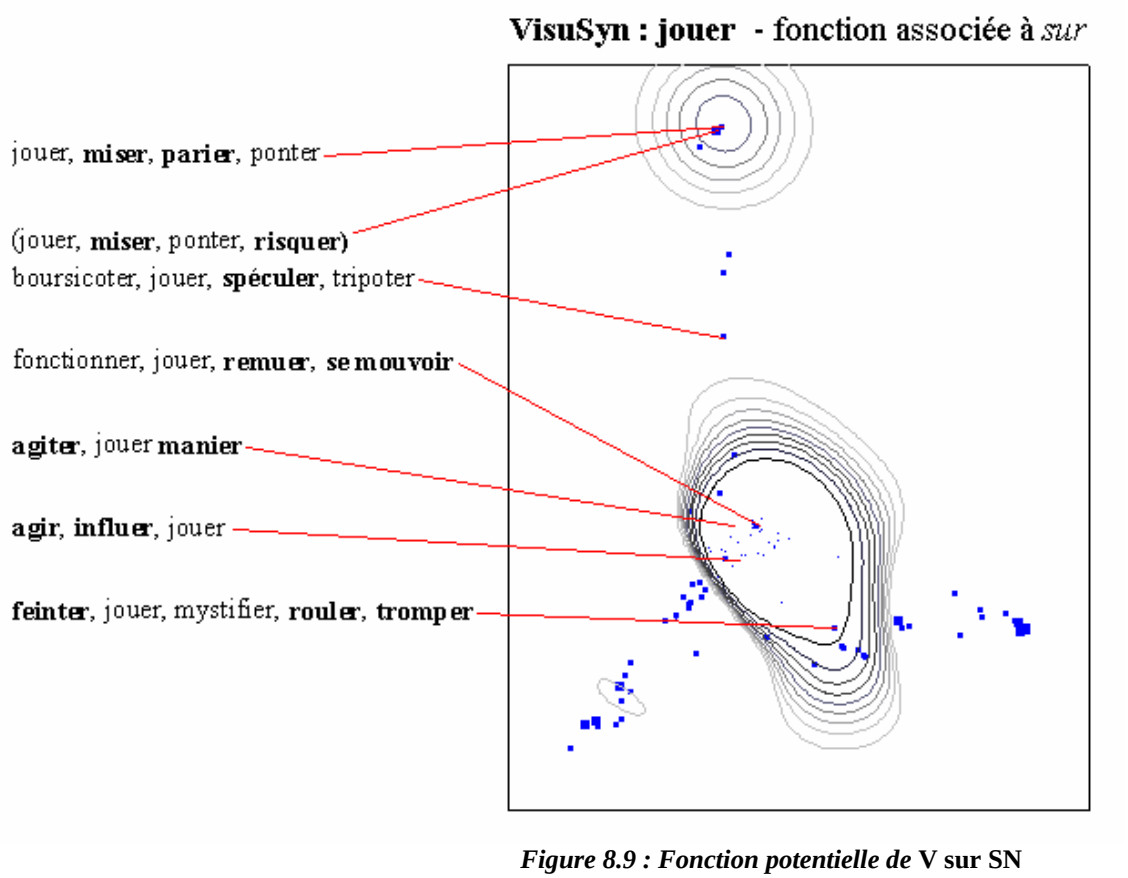


***Figure 8.9 : Fonction potentielle de* V sur SN**

Les cliques présentées sur la figure 8.9 sont celles contenant les synonymes de la dernière colonne. On peut constater que la quasi-totalité des sens proposés par Cadiot se trouve à l'intérieur de la région activée par notre modèle. La seule exception est la représentation de l'énoncé « *jouer sur les grains* » que Cadiot a assimilé à la notion de « *spéculer* ». Nous pensons que dans cet énoncé, on peut aussi assimiler le sens de *jouer* à « *miser, parier* », ce qui correspondrait alors mieux à la région que nous avons calculée. Cependant, même si l'on admet cela il n'en reste pas moins que dans cet énoncé, *jouer* doit être assimilé à *spéculer*. Il y a donc ici une lacune de notre modèle. La principale raison est, selon nous, le fait qu'il y ait peu de cliques dans cette partie de l'espace, ce qui fait que les distances calculées sont accentuées, puisque non contraintes par d'autres cliques ayant un sens différent.

Une autre observation qui conforte les résultats du modèle est que l'ensemble des énoncés proposés par Cadiot n'est pas regroupé dans une zone restreinte de la région activée. Au contraire, les différents sens proposés sont placés aux contours de la région calculée. Par conséquent, la zone délimitée par le modèle possède des caractéristiques qui semblent correspondre à la région activée par la construction *V sur SN*.

8.3.2 La construction prépositionnelle V avec SN

Le mode de présentation et d'évaluation est le même que pour la construction *V sur SN*. Les énoncés proposés par Cadiot sont :

Enoncés illustratifs	Précision du sens	Synonymes correspondants
jouer avec sa santé	*exposer avec légèreté*	*compromettre, risquer*
jouer avec le feu	*manier par défi ou distraction*	*manier, tripoter*
Jouer avec le chien	*tirer un amusement de*	*s'amuser, se divertir*

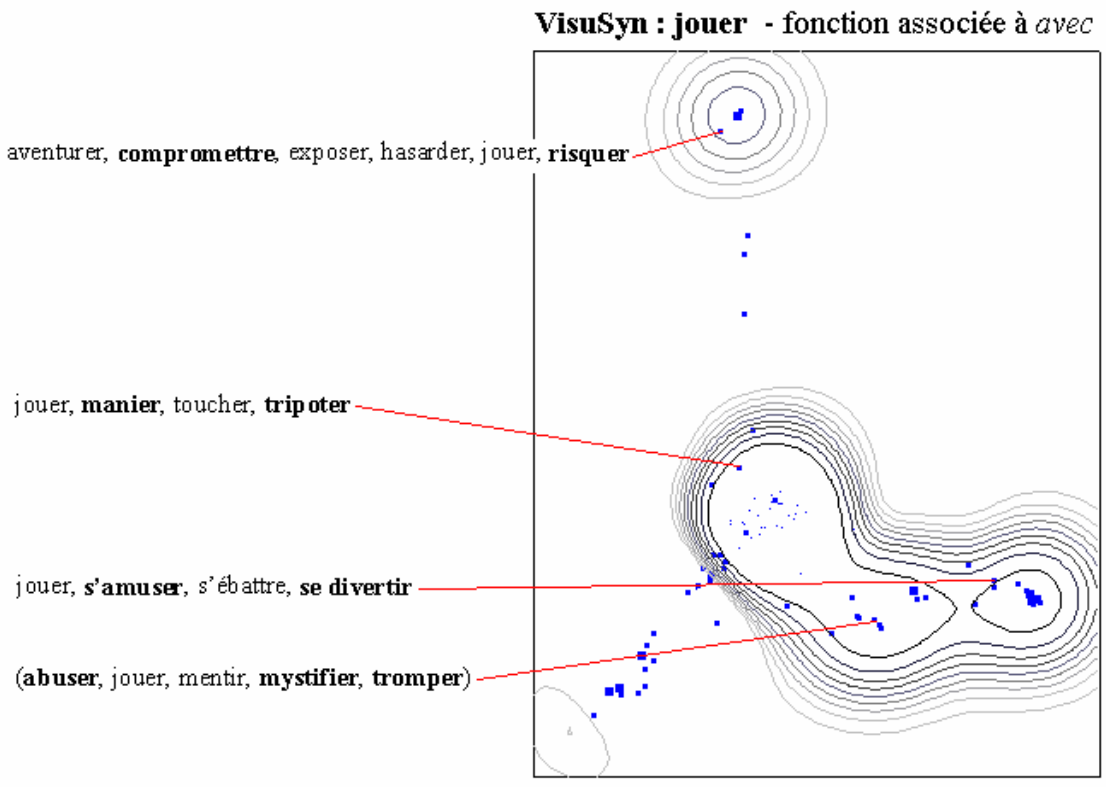


***Figure 8.10 : Fonction potentielle de* V sur SN**

La figure 8.10 met en évidence une superposition encore assez bonne puisque la région activée englobe l'ensemble des énoncés-types, et ces derniers sont assez bien répartis dans cette région activée. On peut même constater qu'une zone de sens a été activée par notre modèle, mais ne possède pas d'équivalent dans les énoncés-types de Cadiot. Cette zone correspond à *jouer* au sens de « *abuser ; mentir ; mystifier ; tromper* ». Un énoncé correspondant pourrait être :

(1) Il joue avec ses employés

On retrouve dans cet énoncé, d'une part le sens de *s'amuser*, mais aussi le sens de *abuser, tromper*.

8.3.3 *Les constructions* V à SN *et* V de SN

Une analyse approfondie a été faite pour chacune de ces constructions (Jacquet, 2002), nous en présentons les fonctions potentielles à titre indicatif (figure 8.11). Dans cette étude, nous nous intéresserons plutôt à une mise en parallèle de ces deux constructions qui toutes deux emploient une préposition appelée parfois « incolore » (Vandeloise, 1993), du fait de l'absence apparente de sens intrinsèque.

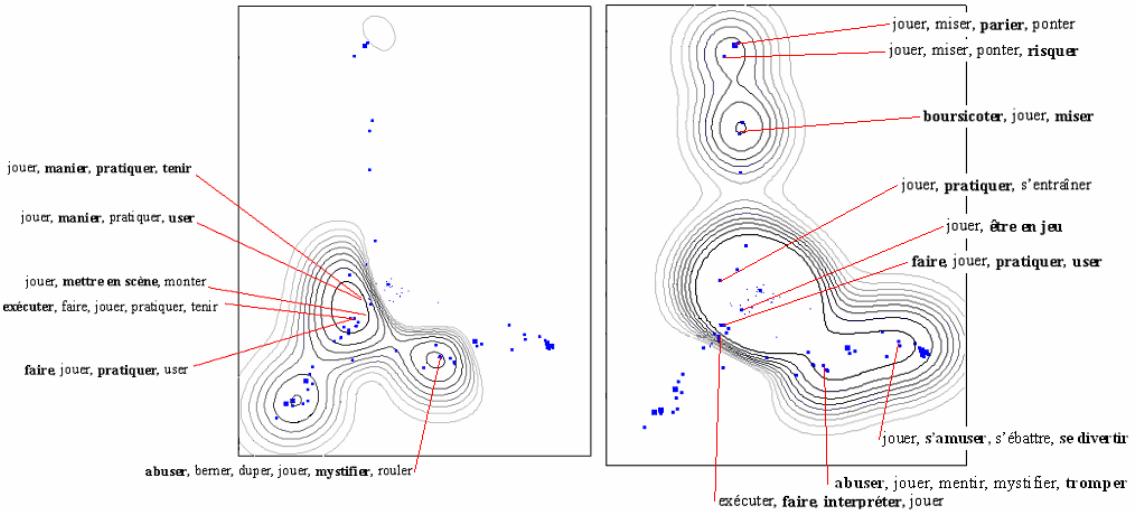


***Figure 8.11 : Fonctions potentielles de* V de SN *et* V à SN**

Au cours de nos manipulations de corpus, nous nous sommes aperçu qu'il était très rare d'observer pour une même construction *V préposition SN*, des cas avec la préposition *à* aux côtés de cas avec la préposition *de*. Il semble donc que leur emploi soit assez hermétiquement séparé pour un verbe donné. Par exemple, pour le verbe *jouer*, *on joue* ***d'un*** *instrument de musique*, alors *qu'on joue* ***à un*** *jeu ou un sport* (Cadiot, 1993, 1999). Et l'inverse est très rare. Ce n'est donc pas un hasard si les espaces sémantiques obtenus pour *V de SN* et *V à SN* sont disjoints.

Nous proposons d'interpréter cette opposition par les constatations suivantes : Lorsque le verbe *jouer* est dans une construction *V de SN*, le *SN* correspond souvent à l'espace de l'activité de jeu, alors que dans une construction *V à SN*, le *SN* correspond à

l'activité en elle-même. Ce qui fait que *à* implique la connaissance préalable de l'espace d'activité. Pour pouvoir prononcer un énoncé tel que *jouer au foot*, il faut au préalable connaître l'espace d'activité auquel cela correspond. Alors que, même si on ne sait pas se servir d'un instrument de musique, étant donné que son espace d'activité est fixé (l'instrument de musique), on peut dire que *l'on joue d'un instrument*[46].

Afin de mieux comprendre cette opposition, nous allons nous intéresser aux cas particuliers, c'est-à-dire les constructions « *jouer préposition SN* » qui acceptent les prépositions *à* et *de* (les énoncés sont tirés de Frantext) :

Jouer au couteau

Jouer du couteau

Le sens de *jouer au couteau* reste indéterminé si cette activité n'est pas définie dans le contexte. C'est-à-dire que pour comprendre le sens de *jouer au couteau*, il nous faut des éléments permettant de construire l'espace d'activité, puisque l'on parle directement ici de l'activité. Alors que *jouer du couteau* fournit directement un sens, c'est-à-dire qu'il fait appel à ce que nous avons comme données préalables sur cet outil, pour construire, dans cette situation, l'espace d'activité.

Jouer au piano

Jouer du piano

Ces deux énoncés sont intéressants parce que le sens du verbe ne change pas : *interpréter, pratiquer*. Or, si l'on reporte ces synonymes dans l'espace sémantique de *jouer*, la zone correspondante est exactement à la frontière entre les régions activées respectivement par les constructions *V de SN* et *V à SN* (cf. figure 8.11). On peut expliquer cela par le fait que la notion « *d'interpréter, représenter, mettre en scène* », véhiculée par « *jouer au piano* » est très proche de « *pratiquer, user, faire* » qui caractérise la notion générale de *jouer d'un instrument*.

[46] Voir Jacquet (2002) pour une description détaillée.

Il jouait du Chopin

Nous jouions à Chopin et Georges Sand

La distinction nette du sens de *jouer* dans ces deux énoncés peut être expliquée par le fait suivant : dans l'énoncé *il joue du Chopin*, quel que soit le nom employé, nous allons construire, avec ce que nous connaissons du nom *Chopin*, un espace d'activité possible pour que cet énoncé soit « pertinent ». Au contraire, on ne peut pas dire, *il joue à Chopin*, sans savoir de quel jeu il s'agit au préalable, puisque cela ne réfère à aucune activité connue hors contexte. La seule possibilité de construction du sens, dans ce cas, est l'identification à l'objet. C'est-à-dire considérer *jouer à* comme *mimer Chopin*, *représenter Chopin*. Il semble que ce ne soit pas propre à *Chopin* ou aux noms propres puisque nous pourrions faire le même raisonnement avec un autre nom :

jouer à la lampe

jouer de la lampe

En admettant qu'il n'existe pas de jeu dénommé *« la lampe »,* la seule interprétation possible de *jouer à la lampe* serait de s'identifier à la lampe[47]. *Jouer de la lampe* correspondrait à jouer d'une manière ou d'une autre en utilisant la lampe comme instrument[48].

Ainsi, l'opposition mise en évidence par les zones de sens calculées semble bien refléter l'opposition linguistique existant entre le sens de *jouer* dans *« jouer **de** quelque chose* » et dans « *jouer **à** quelque chose* ». Le caractère très faiblement sémantique de

---

[47] A titre d'illustration, voici deux extraits renvoyés par Google pour la requête « jouer à la lampe » :

*Tu peux **jouer à la lampe** : il suffit d'avoir des idées très souvent, très souvent.*

*En effet s'il fait nuit, l'écran n'étant pas rétro éclairé, il faut **jouer à la lampe** torche.*

[48] A titre d'illustration, voici deux extraits renvoyés par Google pour la requête « jouer de la lampe » :

*A chaque nouveau lieu, on se surprend à **jouer de la lampe**-torche dans toutes les directions afin de mettre en lumière le moindre mouvement suspect.*

*En tous cas, faites bien gaffe en vous couchant le soir de bien fermer la porte à clé car ce jeu est idéal pour entraîner les cambrioleurs à **jouer de la lampe** torche.*

ces prépositions nous conforte dans l'idée que l'influence calculée ici est celle des constructions syntaxiques, et non celle des prépositions en tant qu'unité grammaticale.

8.3.4 Discussion des résultats

L'évaluation de notre modèle, même imparfaite, a fait ressortir une assez bonne efficacité de notre système à calculer le sens du verbe *jouer* en fonction de la construction prépositionnelle dans laquelle il est employé. Les résultats obtenus semblent confirmer qu'il est possible de réduire automatiquement l'ambiguïté de sens d'un verbe, grâce à sa construction prépositionnelle. Reste maintenant à étendre cette étude sur plusieurs verbes et surtout un plus grand nombre de constructions.

8.4. Evaluation psycholinguistique

L'évaluation que nous venons de présenter a l'inconvénient de ne pas pouvoir être généralisable puisque tous les verbes n'ont pas fait l'objet d'étude aussi détaillée que celle de *jouer* faite par Cadiot. L'autre inconvénient est de se baser sur une seule étude, ce qui implique une part de subjectivité.

Afin de dépasser ces inconvénients, nous proposons une seconde évaluation basée, cette fois-ci sur une étude psycholinguistique. Les atouts de généralisation et de relative objectivation de ce type d'évaluation sont indissociables d'inconvénients tels que le manque de précision, ou encore l'écart inévitable entre les représentations sémantiques des évaluateurs et les représentations sémantiques présentes dans notre corpus d'étude. C'est pour cette raison que nous avons tenu à conserver les deux évaluations afin que l'une comble les lacunes de l'autre.

L'évaluation psycholinguistique que nous allons présenter porte sur nos trois verbes vedettes (*jouer, compter, monter*) pour lesquels nous allons étudier les principales constructions.

- *Jouer : jouer à /jouer sur /jouer avec /jouer transitif direct/jouer intransitif*
- *Monter : monter transitif/monter intransitif*
- *Compter : compter sur/compter transitif*

L'objectif est d'évaluer la pertinence de nos calculs par rapport aux résultats donnés par des locuteurs du français. Il a donc fallu concevoir une tâche de désambiguïsation réalisable à la fois par les sujets et par notre logiciel. C'est ce que nous allons présenter maintenant.

8.4.1 Taux d'adéquation[49]

Pour commencer, il a fallu mettre en place un banc d'essai sur lequel on puisse comparer les résultats calculés automatiquement avec ceux proposés par les évaluateurs. L'idée a été d'utiliser les fonctions potentielles des synonymes, c'est-à-dire de calculer le taux d'adéquation entre la fonction potentielle d'une construction *c* et la fonction potentielle de chaque synonyme du verbe. L'évaluateur devra de son côté avoir une tâche permettant de noter la proximité sémantique de chaque synonyme avec la construction évaluée *c*. Il n'était pas envisageable de travailler avec les listes complètes de synonymes, ce qui aurait rendu la tâche d'évaluation beaucoup trop lourde, nous avons donc décidé pour chaque verbe de sélectionner un nombre restreint de synonymes caractérisant les différents sens que ce verbe peut prendre. Nous avons veillé à ce que ces synonymes correspondent à des cliques bien réparties sur l'espace sémantique :

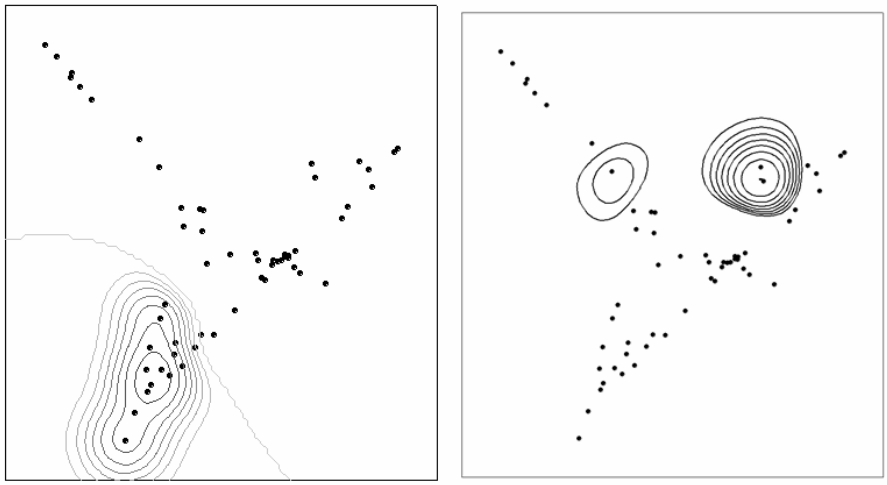
- ***compter** : contenir, espérer, importer, considérer*
- ***jouer** : miser, s'amuser, imiter, influer, se mouvoir, interpréter, abuser, pratiquer*
- ***monter** : manigancer, réaliser, augmenter, hisser, s'élever, gravir*

soit $f_j(x,y)$ et $g_i(x,y)$ les fonctions potentielles respectives du synonyme s_j et de la construction c_i. Le taux d'adéquation entre le synonyme s_j et la construction c_i vaudra :

$$T_{ij} = \frac{\iint f_j \; g_i}{\left(\iint f_j^2\right)^{\frac{1}{2}} \left(\iint g_i^2\right)^{\frac{1}{2}}}$$

[49] Ce taux d'adéquation a déjà été utilisé par Venant (2004) pour une étude sur l'adjectif *sec*.

Plus les zones de sens du verbe et de la construction seront proches, plus le taux sera proche de 1. Reprenons le cas du verbe *compter* dans la construction *V sur SN.* Le taux d'adéquation de *contenir* avec la construction *V sur SN* est égal à 0 (on peut voir les fonctions potentielles respectives dans la figure 8.12). Effectivement *compter sur quelque chose* n'est absolument pas synonyme de *contenir sur quelque chose,* alors que pour *tabler sur quelque chose* on trouve un taux d'adéquation de 0.76, preuve que *tabler* est un bon synonyme de *compter* dans la construction *V sur SN.* Pour chaque verbe étudié (*jouer, compter, monter*) on calculera le taux d'adéquation de chacune des constructions retenues avec chacun des synonymes sélectionnés. (Voir résultats en figure 8.13)



contenir *V sur SN*

Figure 8.12 : Exemple de taux d'adéquation entre un synonyme et une construction égal à zéro

8.4.2 Résultats

Pour tester la validité de ces calculs, nous les confrontons avec les résultats de l'expérience psycholinguistique : on demande à des locuteurs du français de réaliser le même travail que notre modèle : sélectionner parmi les synonymes proposés, celui ou ceux qui décrivent le mieux le sens du verbe en présence d'une construction donnée. Cette expérience a impliqué Laure Sarda pour un contrôle linguistique des énoncés,

Yves Bestgen pour des conseils de méthodologie, ainsi que Jennifer Mercier pour la distribution et saisie des questionnaires. Cette étude a été réalisée sur 112 individus.

La tâche de chaque évaluateur consiste à traiter 25 énoncés. Pour chaque énoncé (extrait aléatoirement du corpus LM3, puis contrôlé comme correspondant bien à la construction recherchée), l'évaluateur doit donner une note à chaque synonyme, en fonction de sa capacité à remplacer le verbe étudié :

1 : pas synonyme

2 : peu synonyme

3 : assez synonyme

4 : très synonyme

Extrait complet	**manigancer** *au sens de* préparer, tramer, combiner	**réaliser** *au sens de* créer, fabriquer, faire	**augmenter** *au sens de* hausser, majorer	**hisser** *au sens de* lever, soulever	**s'élever** *au sens de* atteindre, arriver	**gravir** *au sens de* ascensionner escalader
Les élèves ont monté un coup à leur professeur	4	2	1	1	1	1

Le tableau ci-dessus montre un extrait du questionnaire distribué aux évaluateurs (cf. annexe E pour un extrait plus complet). On peut voir que l'on spécifie le sens de chaque synonyme à l'aide d'un petit ensemble de verbes, afin de contrebalancer le fait que ces synonymes peuvent aussi être polysémiques.

Les résultats sont présentés en page suivante (figure 8.13) où l'on a reporté chaque verbe l'arrondi de la moyenne des notes données par les sujets pour chaque construction. Afin de comparer plus facilement les résultats de notre modèle à ceux de l'expérience psycholinguistique, on leur donne une forme semblable :

Pour l'expérience psycholinguistique, on attribue une note aux synonymes en fonction de leur taux d'adéquation avec la construction : 4 pour une moyenne supérieure à 0.7, 3 pour une moyenne entre 0.5 et 0.7, 2 pour une moyenne entre 0.3 et 0.5 et 1 pour une moyenne inférieure à 0.3. Pour le modèle, on procède de la même manière avec les taux d'adéquation calculés.

On attribue ensuite un taux de réussite permettant d'évaluer la performance du modèle. Si la note est la même pour le modèle et l'évaluation, la réussite est totale (100%). On considère que 3 et 4 sont des notes proches puisqu'elles signifient que le verbe considéré peut être utilisé comme synonyme du verbe étudié, plus ou moins adéquat, dans la construction considérée (80%). De même 1 et 2 sont liées et signifient que le verbe est rejeté, plus ou moins vivement, en tant que synonyme du verbe étudié (80%). En revanche 3 et 2 ne sont que faiblement liées par la notion de changement de sens (20%).

Le tableau ci dessous récapitule les règles d'attribution du taux de réussite.

Note sujet	1				2				3				4			
Note Visusyn	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
% de réussite	100	80	0	0	80	100	20	0	0	20	100	80	0	0	80	100

Verbes étudiés	constructions	Verbes synonymes	Score du modèle	Score des sujets	réussite (%)
compter	SP (sur SN)	énumérer	1	1	100
		contenir	1	1	100
		espérer	4	4	100
		considérer	1	2	80
				moyenne	95
	transitive	énumérer	4	4	100
		contenir	4	2	0
		espérer	1	1	100
		considérer	1	1	100
				moyenne	75
jouer	SP (sur SN)	miser	4	4	100
		s'amuser	1	1	100
		imiter	2	1	80
		interpréter	1	2	80
		abuser	1	1	100
		pratiquer	1	2	80
				moyenne	90
	SP (avec SN)	miser	1	4	0
		s'amuser	4	4	100
		imiter	1	1	100
		interpréter	2	2	100
		abuser	1	2	80
		pratiquer	4	3	80
				moyenne	76.67
	SP (à SN)	miser	1	1	100
		s'amuser	2	4	0
		imiter	1	3	0
		interpréter	4	4	100
		abuser	1	1	100
		pratiquer	4	4	100
				moyenne	66.67
	transitive	miser	2	3	20
		s'amuser	1	1	100
		imiter	4	3	80
		interpréter	4	4	100
		abuser	1	1	100
		pratiquer	4	1	0
				moyenne	66.67
monter	transitive	manigancer	4	3	80
		réaliser	1	4	0
		augmenter	3	1	0
		hisser	1	1	100
		s'élever	2	1	80
				moyenne	52
	intransitive	manigancer	4	1	0
		réaliser	2	1	80
		augmenter	4	4	100
		hisser	1	2	80
		s'élever	3	3	100
				moyenne	72
				moyenne ttle	74.25

Figure 8.13 : Résultats de l'évaluation psycholinguistique

Le taux d'adéquation entre notre modèle et les sujets est de 74.25 %, ce qui est très encourageant compte tenu du fait que le calcul n'est fait qu'à partir de l'influence des constructions. Pour le verbe *compter*, on retrouve dans la figure 8.13 les traces de la fonction potentielle calculée précédemment. En effet, pour la construction *compter SP (sur SP)*, seul le verbe *espérer* possède la note 4, et tous les autres sont à 1. Pour la construction *compter transitif*, les synonymes *énumérer* et *contenir* prennent la note 4.

Les taux d'adéquation sont de 62 % pour le verbe *monter*, 75% pour *jouer*, et 85 % pour *compter*. Plutôt qu'une disparité de réussite selon les verbes, il semble que ces écarts correspondent à des disparités entre constructions. En effet, la moyenne des notes des constructions transitives est de 64.56 %, alors que la moyenne des constructions prépositionnelles est de 82.09 %. Il semble donc que les constructions prépositionnelles soient plus efficaces pour la désambiguïsation.

Parmi les erreurs on peut en noter deux intéressantes. La première est que pour la construction *monter intransitif*, notre modèle donne 4 au verbe *manigancer*, alors que les sujets donnent en moyenne 1. Afin de contrôler ce qui semblait être une erreur de notre modèle, nous avons vérifié dans notre corpus si *manigancer* pouvait effectivement s'employer de manière intransitive. Or, les énoncés correspondant sont principalement des formes interrogatives (*qu'est ce que tu manigances ?)* qui sont considérée à tort comme des constructions intransitives par notre analyseur syntaxique. Ce type d'erreur pourrait être neutralisé en optimisant l'analyseur syntaxique, ou bien en procédant à un filtrage dans notre modèle concernant les formes interrogatives.

La seconde erreur intéressante concerne la construction *jouer transitif*. Notre modèle donne 4 au verbe *pratiquer*, alors que les sujets donnent en moyenne 1. C'est effectivement une erreur de notre modèle. Elle s'explique par le fait que l'on dit *pratiquer la guitare*, mais *jouer de la guitare*. Autrement dit nous sommes dans le cas ou deux verbes sont synonymes mais avec une construction différente. Nous développerons ce point dans le chapitre 10 (§ 10.2.1).

8.5. Combinaison lexique-syntaxe

Maintenant que nous pouvons prendre en compte l'influence des co-textes lexicaux et des co-textes syntaxiques, il apparaît nécessaire ou du moins intéressant d'en combiner les résultats. Nous n'avons pas encore évalué l'efficacité de cette combinaison, néanmoins nous tenons à faire part du type de résultat obtenu.

Prenons les deux énoncés qui suivent comme exemple :

(1) elle joue la fille sérieuse

(2) elle joue avec sa fille

Ces deux énoncés possèdent la même tête de complément (*fille*) et pourtant dans l'énoncé (1) le sens de *jouer* est proche de *interpréter, incarner* alors que dans l'énoncé (2) il est proche de *s'amuser, plaisanter*.

La figure 8.14 rappelle la structure de l'espace sémantique de *jouer* avec les trois axes de sens que nous avons déjà présentée : le premier est *miser, risquer, parier* ; le second est *s'amuser, plaisanter* ; le troisième est *imiter, copier, simuler*. La fonction potentielle du co-texte *fille* sur cette espace (cf. figure 7.15) met en évidence cinq bassins de sens. Quatre d'entre eux correspondent à des sens possibles du verbe *jouer* en présence du co-texte *fille* :

Jouer, rire, s'amuser, se divertir (*Marie joue avec sa fille*)

Figurer, incarner, jouer, reproduire, représenter (*Marie joue la fille sérieuse*)

Jouer, manier, toucher, tripoter (*Ce playboy passe son temps à jouer avec les filles*)

Aventurer, compromettre, exposer, hasarder, jouer, risquer (*Cette fille joue avec sa vie*)

Le cinquième bassin est une erreur. Il correspond à *boursicoter, jouer, spéculer, tripoter* et est activé par la présence de *tripoter*, et de *spéculer* (*spéculer* n'est activé que par une occurrence où *fille* n'est pas rattaché à *jouer* mais à *sentiment* : « Ainsi en peu

de temps tu auras toutes ses breloques, ajouta-t-il en se frottant les mains, heureux de pouvoir **spéculer** sur le sentiment de sa **fille**. » (Balzac H. de, Eugénie Randet, 1843, page 1173)). Autrement dit, même si notre modèle n'avait pas activé par erreur la région *boursicoter, jouer, spéculer, tripoter,* l'ambiguïté du sens de *jouer* est encore très importante puisque les quatre autres bassins correspondent à des sens très différents.

La présence du co-texte *fille* n'est donc pas suffisante pour désambiguïser le verbe *jouer*. Si maintenant on combine ce co-texte avec la construction syntaxique, l'ambiguïté est nettement réduite (cf. figure 8.16 et 8.17) : pour la combinaison *fille + V SN* (*elle joue la fille sérieuse*) la fonction potentielle n'active plus que les régions *« copier, imiter, jouer, mimer, reproduire, simuler »* et *« aventurer, compromettre, exposer, jouer risquer »*, alors que pour la combinaison *fille + V avec SN* (*elle joue avec sa fille*), seule la région « *batifoler, folâtrer, jouer, plaisanter, s'amuser »* est activée.

Cette combinaison est simple puisqu'elle consiste à additionner les fonctions potentielles des différents co-textes et à normaliser la fonction obtenue. Il est important de préciser qu'il s'agit d'une combinaison de deux fonctions potentielles et non une seule fonction correspondant au co-texte « *fille* en tant que tête de complément d'objet direct ». Cela permet d'une part de respecter le cadre théorique que nous nous sommes fixé et d'autre part de calculer nos degrés d'affinités à partir de fréquences plus élevées, donc plus fiables. On pourrait envisager une combinaison plus élaborée, dépendant du verbe à désambiguïser ou du degré de fiabilité des fonctions potentielles obtenus, ou encore dépendant de la capacité des unités co-textuelles à désambiguïser, etc. Nous reviendrons sur ce point dans le chapitre 10 (§10.1.2).

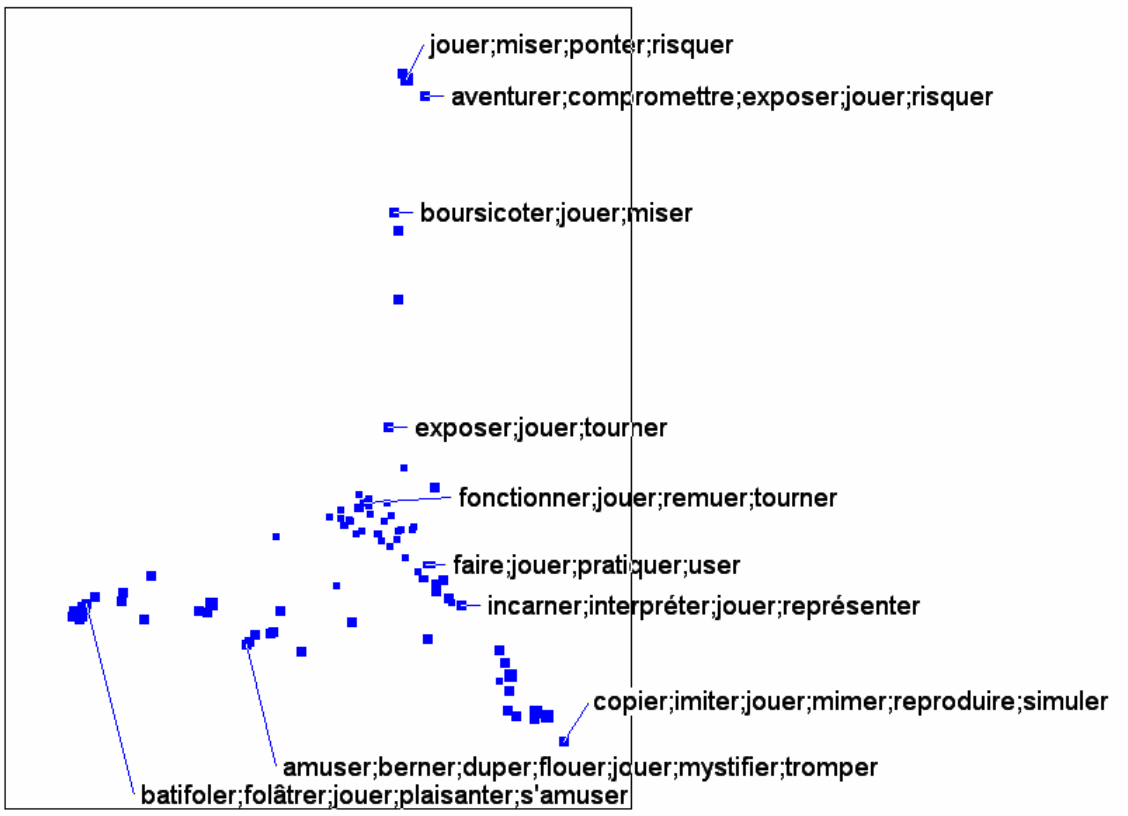


Figure 8.14 : Espace sémantique du verbe* jouer *(rappel)

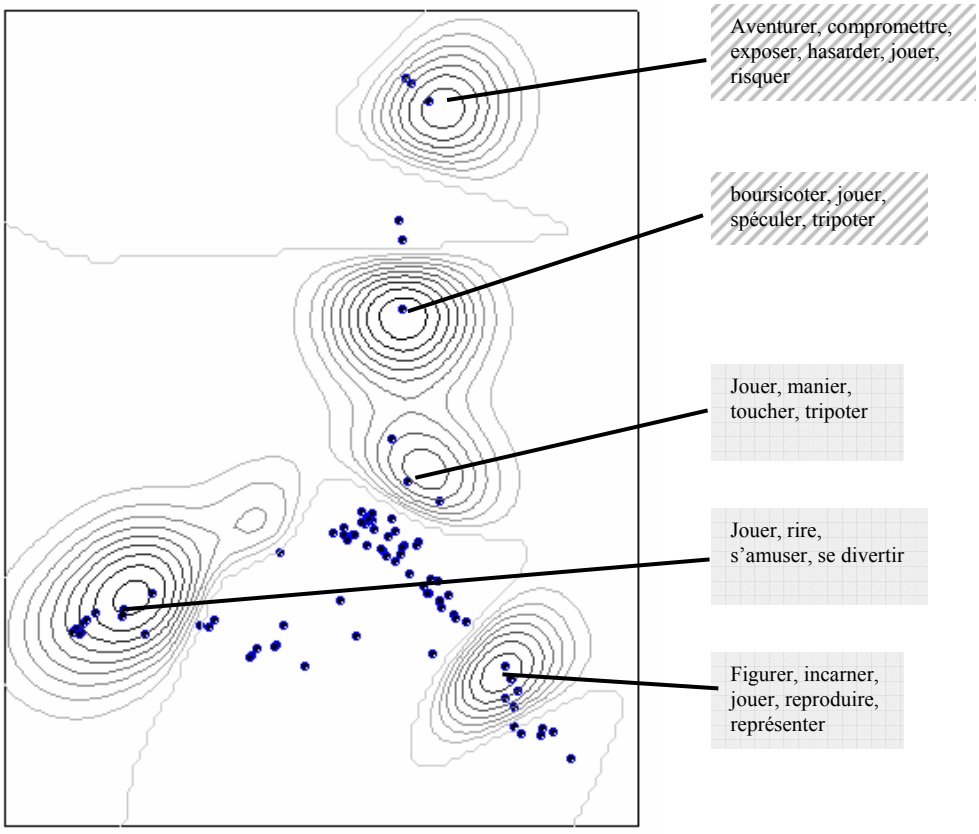


***Figure 8.15 : Fonction potentielle du co-texte* fille**

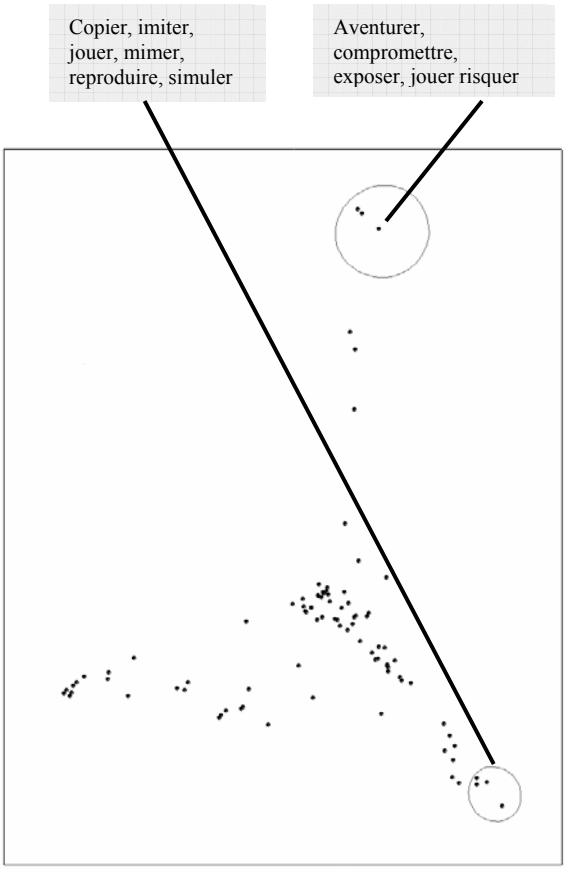


***Figure 8.16 : Combinaison des fonctions potentielles du co-texte* fille *et de la construction* V SN.**

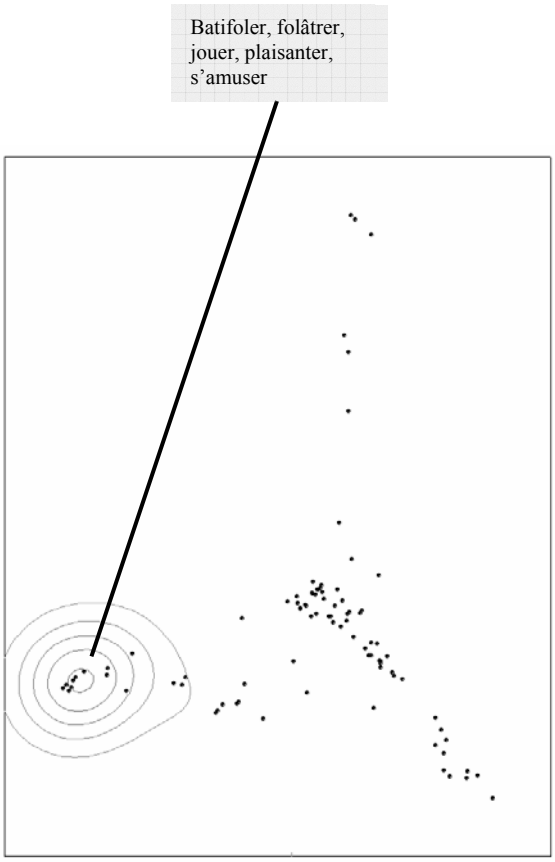


***Figure 8.17 : Combinaison des fonctions potentielles du co-texte* fille *et de la construction* V avec SN.**

8.6. Discussion

A l'aide des deux modes d'évaluation que nous venons de présenter, nous pouvons dire que notre modèle est capable de réduire l'ambiguïté d'un verbe polysémique uniquement à l'aide de la construction syntaxique dans laquelle il est employé. Nous pouvons aussi apporter une preuve supplémentaire au fait que les constructions syntaxiques sont des unités linguistiques porteuses d'un sens intrinsèque et que ce sens permet de réduire l'ambiguïté d'un verbe. Et ce, même pour des constructions très simples comme la construction transitive. Les premiers résultats obtenus sur la combinaison du co-texte lexical et syntaxique sont encourageants mais encore limités. L'étude de cette combinaison nécessiterait une évaluation et notamment mériterait d'être comparée aux résultats obtenus avec une approche classique, c'est-à-dire sans dissocier syntaxe et lexique.

9. Les constructions verbales : améliorations du modèle

Des résultats que nous venons de décrire émergent deux problèmes importants qu'il convient de résoudre pour pouvoir aller plus loin. Le premier concerne la représentation des fonctions potentielles d'une construction verbale. Nous avons pu observer, notamment avec la fonction potentielle de *V de SN* sur le verbe *jouer* ou encore la fonction potentielle de *V sur SN* sur le verbe *compter*, que certaines régions étaient activées à tort pour une raison parfaitement explicable, qui ne dépend pas de notre méthode mais des résultats de l'analyse syntaxique en amont. L'analyseur syntaxique que nous utilisons, Syntex, construit des relations de dépendance. On sait que l'une des principales difficultés auxquelles se heurtent ces analyseurs concerne les questions de rattachements prépositionnels (Bourigault et Fabre 2000, Frérot 2003, Bourigault et Frérot 2004). Considérons les exemples suivants :

(1) *Les variations soumettent les particules* ***à des mouvements*** *...*

(2) *Les variations soumettent les particules* ***du milieu*** *...*

Ces deux énoncés ont la même construction de surface (S V SN SP), cependant, dans (1) le SP se rattache au verbe (*soumettre à mouvements*) alors que dans (2) le SP se rattache au nom (*particules du milieu*). Les erreurs de rattachement prépositionnel sont bien sûr préjudiciables pour notre système, puisqu'elles introduisent du bruit (si un complément est rattaché à tort au verbe) et du silence (si à l'inverse un complément du verbe est rattaché à un nom). Cependant, une analyse rapide de nos résultats montre que ce n'est pas le problème le plus important.

En revanche, ce n'est pas à proprement parler des erreurs d'analyse, mais plutôt une insuffisance de l'analyseur qui provoque le problème le plus grave : il ne fait pas la

distinction entre compléments essentiels et compléments circonstanciels. Par exemple, considérons, les quatre expressions suivantes :

(3) *passer le sel à son voisin*

(4) *passer à la télévision*

(5) *passer ses vacances à la mer*

(6) *passer un accord commercial avec ses concurrents à Paris*

Dans les quatre cas, le syntagme prépositionnel introduit par *à* sera traité de la même manière, simplement comme un complément du verbe *passer.* Or il y a clairement de grandes différences entre eux : dans (3) et dans (4), *à son voisin* et *à la télévision* sont des compléments essentiels du verbe, partie intégrante de la construction, indispensables à la compréhension du sens que prend *passer* dans ce co-texte. Tandis que dans (6) *à Paris* est un complément circonstanciel, accessoire, qui ne joue aucun rôle pour le sens du verbe. Le cas (5) est intermédiaire : *à la mer* est un complément circonstanciel, mais plus « intégré » à la construction. On peut le supprimer, mais il faut alors le remplacer par un autre complément, par exemple de manière : *passer ses vacances en famille*.

Le fait de ne pas pouvoir se limiter aux compléments essentiels a de lourdes conséquences quand on calcule l'influence de la construction (cf. chapitre précédent). Ainsi, la région *calculer, énumérer* du verbe *compter* est activée pour la construction *V sur SN* par des énoncés du type *compter/calculer sur ses doigts*. Or dans ces énoncés *sur ces doigts* est un complément circonstanciel de manière, ce qui est très différent de la construction *V sur SN* qui donne à *compter* le sens de *tabler*, *escompter*. De même, la région *imiter, copier, simuler* du verbe *jouer* est activée pour la construction *V de SN* par des énoncés du type *jouer/imiter/copier de façon remarquable.*

Nous avons donc entamé des recherches, qui en sont encore à leurs débuts, pour traiter ce problème. Le principe n'est pas d'essayer de distinguer les compléments essentiels des circonstanciels, mais plutôt de mesurer l'aspect plus ou moins essentiel

d'un complément sur une échelle graduelle, avec l'idée qu'il existe ici aussi un continuum entre ces deux types de compléments (Victorri et Fuchs, 1996 ; Fabre et Frérot, 2002). Nous allons présenter dans la première partie de ce chapitre les premières expérimentations que nous avons réalisées.

Le deuxième problème fondamental à résoudre concerne les calculs de fréquence. Nous avons évalué notre modèle en utilisant presque exclusivement des éléments co-textuels très fréquents dans le corpus. Le cas de *monter* avec le co-texte *diamant* est une exception : les fréquences qui ont permis de construire la fonction potentielle de *diamant* sont très faibles puisque comprises entre 0 et 5. Même si les résultats sont corrects dans ce cas précis, les calculs avec des fréquences aussi faibles ne sont pas fiables, et donnent régulièrement des résultats assez mauvais.

Autrement dit, notre modèle peut facilement désambiguïser le sens de *jouer* dans *jouer de la guitare*, mais les résultats sont beaucoup plus incertains pour l'énoncé *jouer du luth*. L'idée est de remplacer une unité co-textuelle rare par une classe de mots qui influent de la même manière sur le verbe étudié, c'est-à-dire, dans notre exemple, remplacer *luth* par la classe {*luth, guitare, piano*}. L'idée d'introduire des classes de mots dans un modèle de désambiguïsation n'est pas nouvelle. Nous présenterons dans la deuxième partie de ce chapitre les recherches que nous avons menées dans ce domaine, en essayant de montrer la singularité de ce que nous avons appelé des *classes de sélection distributionnelle*.

9.1. Degré « d'essentialité » d'un complément

Fabre et Frérot (2002) ont proposé une méthode de repérage automatique en corpus du caractère essentiel ou circonstanciel des groupes prépositionnels. Il s'agit d'une méthode endogène : on mesure des fréquences de cooccurrences sur le corpus qu'on est en train d'analyser. Le principe est de mesurer la productivité de certains

schémas syntaxiques, c'est-à-dire le nombre de mots différents qui peuvent entrer dans un même schéma. Deux mesures sont utilisées :

(a) la productivité d'un verbe donné avec une préposition donnée : combien de noms différents trouve-t-on comme complément du verbe avec cette préposition ? Par exemple la productivité du couple (*alterner, avec*) est de 7 sur leur corpus[50] parce qu'on trouve, en position de complément de *alterner avec*, les 7 noms suivants : *cendre, crue, lit, masse, période, section, surface*.

(b) la productivité d'une préposition donnée et d'un nom donné : combien de verbes différents trouve-t-on comme recteur d'un syntagme prépositionnel constitué par cette préposition et ce nom ? Par exemple la productivité du couple (*à, quaternaire*) est de 3 sur leur corpus parce qu'on trouve *au quaternaire* en position de complément des 3 verbes suivants : *englacer, évoluer, subir*..

En croisant ces deux mesures, les auteurs définissent des critères de classement des compléments : si la productivité (a) est forte et la productivité (b) nulle, alors le complément est classé comme essentiel. Si c'est l'inverse, (a) nulle et (b) forte, il est classé comme circonstanciel. Une évaluation sur 100 triplets de type (verbe, préposition, nom) a donné un taux de bonnes réponses très correct.

Néanmoins, nous n'avons pas utilisé la même méthode pour deux raisons :

- d'une part, nous travaillons sur des verbes polysémiques qui admettent plusieurs constructions. Ce que nous cherchons à faire, c'est différencier les compléments essentiels et circonstanciels pour une même préposition (par exemple distinguer *jouer de la guitare* et *jouer de façon remarquable*). La mesure (a) est donc très peu pertinente, car elle est en général forte, puisqu'il existe des compléments essentiels avec la préposition étudiée. La mesure (b) à elle seule ne permet alors de classer que peu de compléments.

[50] Le corpus sur lequel elles ont mené leurs expérimentations est un corpus de géomorphologie. L'analyseur syntaxique utilisé est Syntex.

- d'autre part, plutôt que de travailler avec un petit corpus spécialisé, nous voulions tester une approche sur de très vastes données textuelles, pour les raisons que nous avons déjà évoquées (cf. introduction et chapitre 5 § 5.3).

Nous avons choisi de travailler directement sur le Web, ce qui, comme nous allons le voir, pose des problèmes méthodologiques particuliers. Mais, comme nous le verrons aussi, la taille des données disponibles permet d'utiliser des tests qui ne sont pas possibles sur des petits corpus.

9.1.1 Utilisation du Web comme « corpus »

L'idée d'utiliser le Web comme outil pour le traitement automatique de la langue est relativement nouvelle. Cela permet d'augmenter de manière très importante les ressources linguistiques disponibles pour toute sorte de tâches. Pour ne donner qu'un exemple, Resnik (1999) a montré qu'avec des méthodes très simples, on pouvait obtenir des corpus parallèles bilingues de taille respectable pour un très grand nombre de langues, alors que jusqu'alors on ne disposait de ressources de grande taille que pour un petit nombre de couples de langues, comme par exemple le français et l'anglais avec le English-French Canadian Hansard. En quelques années, le nombre d'études exploitant directement ou indirectement le Web a augmenté considérablement. Kilgarriff et Grefenstette (2003) proposent une description de ce nouvel outil et de ses utilisations en traitement automatique de la langue. Notamment ils évoquent le fait que dans la majorité des cas, l'accès au Web se fait à partir de moteurs de recherche tels que Google. Ils proposent alors de décrire les caractéristiques des résultats fournis par un moteur de recherche pour une utilisation linguistique du Web :

- Les résultats du moteur de recherche ne donnent pas assez d'instances (1 000 ou 5 000 maximum)
- Ils ne présentent pas assez de contexte pour chaque instance (environ 10 mots pour Google)

- Ils sont sélectionnés en fonction de critères qui sont, pour des linguistes, biaisés (les pages ayant les mots de la requête dans le titre occupent les premières lignes)
- Ils ne permettent pas de spécifier les recherches avec des critères linguistiques tels que la forme du mot, la ponctuation, et encore moins la partie du discours du mot.
- Les statistiques sont peu fiables étant donné que les fréquences données par « pages contenant l'expression x » varient d'un moteur de recherche à l'autre et ne sont que des estimations.

Kilgarriff et Grefenstette incitent néanmoins les linguistes à exploiter cette source d'information, à condition de garder à l'esprit les critiques énoncées. L'intérêt du Web est qu'il contient une quantité colossale de texte, il est pour certaines langues la seule base textuelle existante, il est gratuit, et est disponible instantanément.

On peut noter, parmi les utilisations du Web en TAL, plusieurs travaux traitant de la désambiguïsation automatique. Notamment Agirre et Martinez (2004), dont l'objectif est d'utiliser le Web comme source d'exemples pour les sens des mots. L'idée de départ est que pour un sens de mot donné, si le mot possède pour ce sens un synonyme monosémique, alors les énoncés contenant ce synonyme doivent être très proches du sens du mot visé. Ces énoncés peuvent alors être utilisés pour l'apprentissage d'un classifieur du sens visé du mot étudié. Ils montrent qu'il est possible de rassembler des exemples basés sur ces « monosèmes relatifs » pour quasiment tous les noms de WordNet.

On trouve aussi des études plus proches de la tâche que nous nous sommes fixée, qui s'intéressent notamment au rattachement prépositionnel (Volk, 2001 ; Gala 2003). Le principe est le même que sur corpus. Cela consiste, lorsqu'il y a une indétermination de rattachement prépositionnel, à construire une requête de moteur de recherche pour les deux rattachements possibles : par exemple, si l'on reprend l'énoncé (1) *Les variations soumettent les particules à des mouvements ...* , les deux requêtes seraient

"soumettent à des mouvements" et *"les particules à des mouvements"*. On choisit alors le rattachement qui obtient le meilleur score. Nous allons voir que notre méthode est très proche de ce principe.

9.1.2 Méthode

Le principe de notre méthode consiste à essayer d'évaluer un *degré d'essentialité* d'un complément, puisque nous pensons qu'il s'agit d'un phénomène graduel. Pour cela, nous avons commencé en choisissant deux tests très simples et nous avons construit des requêtes Google censées effectuer ces tests sur chaque couple verbe+complément. Les tests proposés sont les suivants :

Test1 : Est-ce que le couple verbe + complément se retrouve fréquemment ?

Test2 : Est-ce que le complément peut être antéposé ?

Le premier test est très basique. Il est basé sur l'hypothèse qu'un verbe aura plus d'affinité avec ses compléments essentiels qu'avec les circonstanciels. Le deuxième test est linguistiquement plus probant : la possibilité de l'antéposition est un critère fort pour définir le caractère circonstanciel d'un complément. Il est clair que ces deux tests ne sauraient suffire à eux seuls pour obtenir la gradation que nous recherchons. Nous pensons cependant qu'ils peuvent donner une première indication sur l'intérêt de ce type de méthode.

Nous présentons maintenant le mode opératoire que nous proposons pour effectuer ces deux tests.

9.1.3 Test1 : fréquence de la construction verbe + complément

Nous avons basé notre évaluation sur le calcul de l'information mutuelle entre le verbe et le complément. Rappelons que l'information mutuelle $MI(u_1, u_2)$ entre les deux expressions u_1 et u_2 est donnée par :

$$MI(u_1, u_2) = \log \frac{\frac{freq(u_1 + u_2)}{freq(phrases)}}{\frac{freq(u_1) \times freq(u_2)}{freq(phrases)^2}}$$

où $freq(u_1 + u_2)$ est le nombre de cooccurrences de u_1 et u_2 dans la même phrase, $freq(u_1)$ est le nombre d'occurrences de u_1, $freq(u_2)$ est le nombre d'occurrences de u_2 et $freq(phrases)$ est le nombre de phrases du corpus. Ce dernier nombre, dans notre cas est inconnu. Mais comme ce qui va importer, c'est les valeurs relatives des IM les unes par rapport aux autres, il nous suffit que ce nombre soit constant pour toutes les IM que l'on va mesurer. Autrement dit, l'hypothèse que nous faisons, c'est que le nombre de pages indexées par Google est constant. Ceci est évidemment faux sur le long terme, mais si toutes les mesures sont faites dans un intervalle de temps assez court (de l'ordre de la semaine), l'approximation reste très raisonnable.

Prenons l'exemple de la construction *jouer de la guitare.* Pour calculer l'IM entre *jouer* et *de la guitare* nous allons faire trois requêtes : jouer, "de la guitare" et "jouer de la guitare", et considérer que les nombre de pages annoncées « sur un total d'environ X pages» pour ces trois requêtes, N_1, N_2, et N_{12}, mesurent respectivement les fréquences $freq(u_1)$, $freq(u_2)$ et $freq(u_1 + u_2)$. Si l'on appelle C la valeur supposée constante de log(*phrases*), on a :

MI(*jouer*, *de la guitare*) = $\log(N_{12}) - \log(N_1) - \log(N_2) + C$

L'hypothèse que nous faisons est que plus cette information mutuelle est grande, plus le complément a de chance d'être essentiel.

9.1.4 Test2 : antéposition du complément

Le principe de mesure est le même, c'est-à-dire que l'on utilise l'information mutuelle mais cette fois, en sens inverse. C'est-à-dire que plus l'information mutuelle sera importante, plus le complément a de chance d'être un complément circonstanciel. Les requêtes correspondant respectivement aux fréquences freq(u1), freq(u2) et freq(u1 + u2) seront il, "de la guitare" et "de la guitare il". L'idée étant que l'on trouvera moins facilement des énoncés du type *de la guitare il joue/il fait* que des énoncés du type *le matin il joue/il fait quelque chose*. Trois points sont à souligner. D'une part, des énoncés corrects contenant "*de la guitare il*" existent (*de la guitare, il en joue très bien*), mais ces énoncés sont négligeables en termes de fréquence. La deuxième remarque concerne le choix du pronom *il*. Ce choix ne se justifie que par la volonté de formuler une requête correspondant à une situation d'antéposition d'un complément la plus fréquente possible (afin d'avoir des résultats les plus fiables possibles en terme de fréquence). La troisième remarque concerne le fonctionnement du moteur de recherche. Google ne se préoccupe pas de la ponctuation, ainsi l'énoncé "il joue **de la guitare. Il** chante très bien" sera comptabilisé dans la fréquence de *"de la guitare il"*. C'est un problème que nous ne pouvons pas éviter pour l'instant et qui biaisera inévitablement nos résultats.

9.1.5 Premiers résultats

Comme nous l'avons dit, cette série de tests est encore à l'état d'expérimentation. Actuellement nous développons un système en *php* afin d'extraire automatiquement les fréquences fournies par Google, et de calculer l'information mutuelle pour un nombre important de rattachements prépositionnels. La figure 9.1 présente la base du système d'appel en *php*.

```
function($query){
        global $temps;
        global $connection;
        global $cookies;
        global $tempsRequetes;
        $params= array( ie => "iso-8859-1",
                        oe => "iso-8859-1",
                        hl => "fr",
                        num=> "20",
                        meta=> "cr=countryFR",
                        q => $query);
        $temps->startChrono("toto");
        $answer2 = getHttpRequest("www.google.fr","/search",$params,$cookies,$connection,$_,"/temp/answer.http");
        $temps->stopChrono("toto");
        $tempsRequetes += $temps->getChrono("toto");
        $content = $answer2[content];
        if (preg_match("~sur un total d'environ <b>(.*?)</b>~ms",$content,$match)) {
                $nbmatch = intval(str_replace(",","",$match[1]));
        }elseif (preg_match("~sur <b>(.*?)</b>\. Recherche effectu\Ã©e en~ms",$content,$match)) {
                $nbmatch = intval(str_replace(",","",$match[1]));
        }elseif (preg_match("~Aucun document ne correspond aux termes de recherche spÃ©cifiÃ©s ~ms",$content,$match)) {
                $nbmatch = 0;
        }
        return $nbmatch;
}
```

Figure 9.1 :extrait du système de requête automatique en php

compléments	***[expression]***	**"jouer *[expression]*"**	**"*[expression]* il"**	**test 1**	**test2**	**test1 - test2**
sur le canapé	234 000	87	713	0,00	0,00	0,00
de la guitare	1 450 000	144 000	9 600	5,59	0,77	4,82
le matin	2 350 000	310	86 200	-1,04	2,48	-3,52
avec les enfants	767 000	12 700	11 400	3,79	1,58	2,21
à sèbe	215	2	0	3,22	-∞	+∞
aux avions	81 000	575	218	2,95	-0,13	3,08
en public	3 000 000	734	19 200	-0,40	0,73	-1,13
à paris	12 100 000	685	366 000	-1,88	2,29	-4,17
sur la table	2 370 000	265	42 200	-1,20	1,76	-2,96
son ranz	19 400	19 000	876	7,88	2,69	5,19

Figure 9.2 : Tableau récapitulatif des IM calculés

On trouvera figure 9.3 les premiers résultats obtenus avec le verbe *jouer* et quelques uns de ses compléments. Puisque les informations mutuelles ne sont définies qu'à une constante additive près, nous avons arbitrairement fixé à 0 les mesures pour le premier complément testé *sur un canapé*. Rappelons aussi que les résultats que nous présentons ici ne sont là que pour donner une idée de ce que l'on peut attendre de cette évaluation et que seuls les résultats relatifs nous intéressent. La figure 9.2 détaille les fréquences extraites de Google ainsi que les IM calculées pour les deux premiers tests.

Le choix de ces dix compléments se justifie ainsi. *Jouer de la guitare* et *jouer le matin* sont là comme des compléments typiques respectivement de la classe des essentiels et des circonstanciels. C'est-à-dire des compléments qui doivent absolument pouvoir être désambiguïsés. *Jouer aux avions* et *jouer sur le canapé* nous ont semblé correspondre à des usages respectivement de complément essentiel et de complément circonstanciel moins typiques. Nous avons ensuite choisi des compléments qui nous avaient posé problème lors d'une première tentative de distinction automatique des compléments. Ainsi, *avec les enfants, à Paris* et *en public* sont des compléments circonstanciels, mais ces compléments sont très courants, et peuvent donc être confondus avec des compléments essentiels. A l'inverse, les compléments essentiels contenant un mot rare peuvent être confondus avec des compléments circonstanciels, c'est pourquoi nous avons choisi les compléments *à sèbe*[51] et *son ranz*[52].

Pour visualiser les résultats, on a placé les dix compléments sur des axes qui indiquent le degré d'essentialité que nous avons obtenu. On les trouvera figure 9.3[53]. Comme on peut le constater sur la figure, les résultats pour le test 1, représentés en (a), sont assez satisfaisants. Les compléments essentiels (*son ranz*, *de la guitare*, *à sèbe* et *aux avions*) sont bien séparés des compléments circonstanciels, à l'exception de *avec les enfants*, qui se retrouve au milieu des compléments essentiels, mais cette erreur n'est pas très grave, dans la mesure où il s'agit du complément circonstanciel le moins périphérique.

Les résultats sont beaucoup moins bons pour le test 2, représenté en (b). En effet, *son ranz* se trouve classé à l'opposé de là où il devrait être, et *de la guitare* est au milieu des compléments circonstanciels. Les raisons de ces mauvais classements sont à

51 Sèbe est un ancien jeu des cours de récréation (on disait autrefois jouer à sèbe ou au cheval fondu).

52 Ranz des vaches : Air populaire, aux nombreuses variantes, que chantent les bergers suisses.

53 Sur le graphique, la valeur $-\infty$ obtenue pour *sèbe* dans le test 2 est remplacée par une forte valeur négative (-4). D'autre part, l'axe a été inversé pour le test 2 de manière à ce que le gradient circonstanciel → essentiel soit toujours orienté du bas vers le haut.

rechercher avant tout dans l'approximation de nos requêtes Google. Comme nous l'avons noté, Google ne tient pas compte des ponctuations. Dans une quart environ des 9 600 réponses à la requête "*de la guitare il*", *de la guitare* et *il* ne sont pas dans la même phrase. Exemples :

L'un d'eux était Trille Labarre, un virtuose ***de la guitare. Il*** *composa de la musique pour guitare seule, guitare et violon et pour guitare et voix. ...*

Il fit plusieurs suggestions qui conduisirent à des améliorations valables ***de la guitare. Il*** *était un compositeur prolifique.*

Dans pratiquement tous les autres cas, *de la guitare* est complément d'un nom, et pas complément du verbe qui suit *il*. Exemples

Non content d'être juste un génie ***de la guitare, il*** *injecte de la comédie physique d'une vivacité sans pareil dans son jeu.*

Virtuose ***de la guitare, il*** *compose des mélodies simples et accrocheuses aux arrangements délicats.*

Quant aux 876 réponses à la requête "*son ranz il*", elles proviennent d'un très petit nombre de sites, et l'expression ne fait pas partie du texte affiché dans les pages : elle se trouve dans le source html, parmi une liste de mots-clés hétéroclites. Exemple :

```
<dt> <a href="../telecharger_logiciel_emule_gratuitement/"
title="telecharger logiciel emule gratuitement">telecharger
logiciel emule gratuitement</a></dt> <dd>  Vente Aux Enchères
manquez des chansons sentez elle nu le duc tenez son ranz il
plait telecharger logiciel emule gratuitement direz les paul
mesdames.</dd>
```

Il s'agit peut-être du résultat d'une automatisation de l'écriture de ces mots-clés, dont l'objectif serait de tromper les logiciels d'indexation des moteurs de

recherche ?[54] C'est en tout cas pour notre requête un « accident » inhérent à l'utilisation du Web.

Mais quand on examine le troisième axe (c), obtenu en retranchant la deuxième mesure à la première, on a la bonne surprise de constater que le classement est encore meilleur que sur l'axe (a), bien qu'on l'ait combiné avec les mauvais résultats de (b) ! En effet, cette fois-ci, compléments essentiels et circonstanciels sont tous bien placés, et, notamment, l'ordre dans lequel sont placés les circonstanciels est tout à fait correct : les plus périphériques sont à juste titre *à Paris* et *le matin*, suivis de *sur la table, en public* et *sur le canapé*, qui sont effectivement plus « proches » du verbe (*sur la table*, proche de *le matin* sur l'axe, étant d'ailleurs celui des trois qui spécifie le moins d'action de *jouer*), et enfin de *avec les enfants*, dont on a déjà dit qu'il devait effectivement être à la limite des compléments essentiels.

[54] Si notre explication est la bonne, on ne peut que douter de la qualité du système automatisé qui a sélectionné le mot-clé *ranz* comme particulièrement alléchant pour les surfeurs !! A moins qu'il ne s'agisse de « noyer » les mots clés cibles dans un vocabulaire plus anodin pour éviter d'être repérés ?

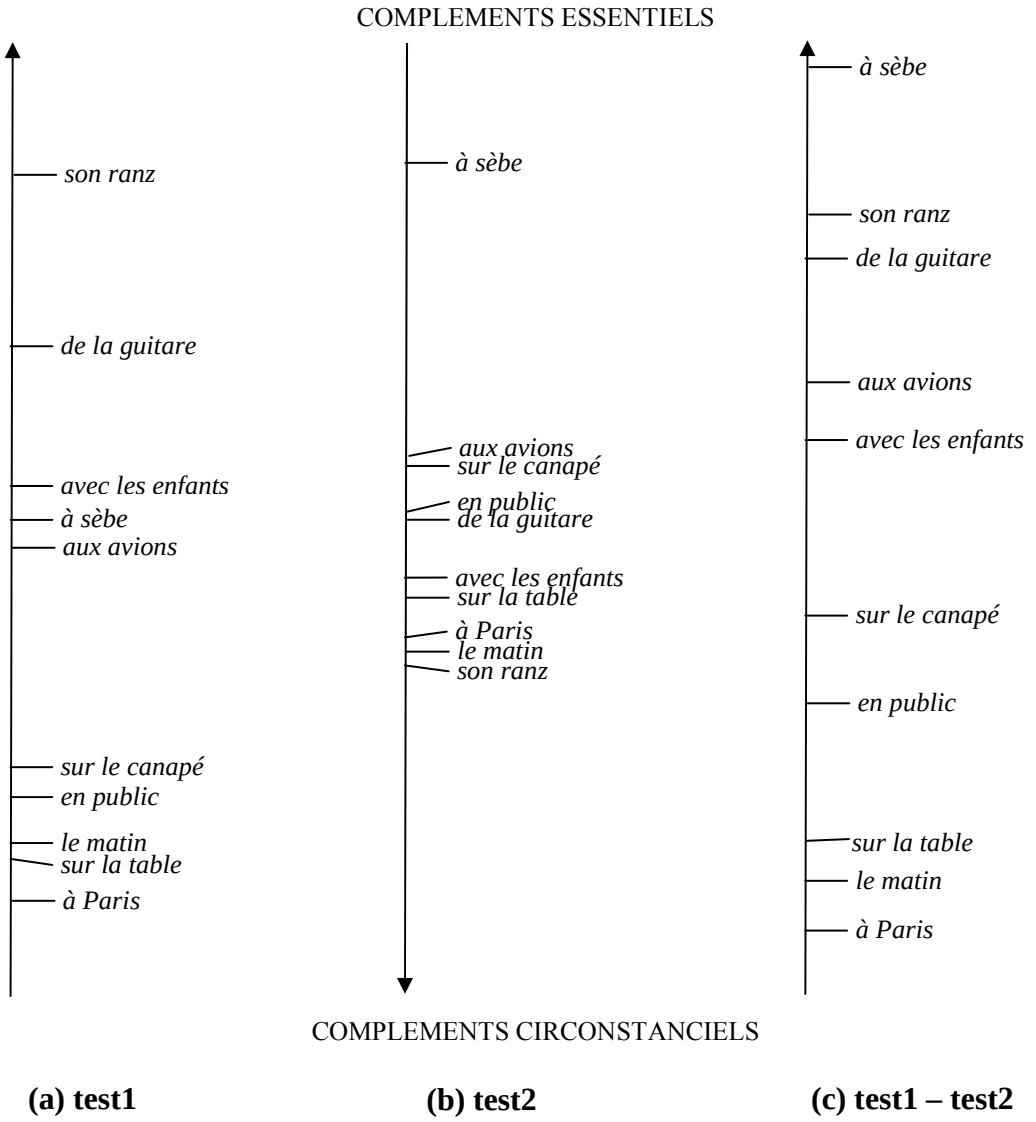


Figure 9.3 : Visualisation des résultats

9.1.6 Discussion

Il serait bien sûr tout à fait prématuré de tirer des conclusions définitives de cette toute petite expérimentation. Néanmoins, ces premiers résultats sont très encourageants, et ils montrent qu'il serait intéressant de poursuivre dans cette direction. En effet, ils montrent que le Web est une ressource intéressante pour cette tâche : malgré les imperfections des requêtes Google et du codage *html* relativement à la tâche visée (cf. la

requête "*son ranz il*"), les résultats sont globalement corrects : le bruit introduit par ces imperfections ne détériore pas trop les performances.

Qui plus est, il semble possible d'améliorer sérieusement la méthode en évitant un certain nombre de ces défauts : par exemple, en éliminant les réponses qui comportent un point au milieu de l'expression recherchée (Si le nombre de réponses est trop important pour faire cela exhaustivement, on peut procéder par sondage pour donner une approximation du nombre de réponses à éliminer). Pour des erreurs plus délicates à traiter (comme les cas où *de la guitare* est complément de nom au lieu de complément du verbe), on pourrait envisager carrément de procéder à une analyse syntaxique avec un analyseur tel que Syntex (là aussi, on pourrait procéder par sondage pour ne pas trop alourdir les temps de calcul).

Enfin, le fait que la combinaison des deux tests donne de meilleurs résultats que chaque test séparé est très prometteur : il existe en effet beaucoup de tests possibles du même genre que l'on pourrait combiner, utilisant ainsi la redondance de ces tests pour pallier les erreurs éventuelles de l'un ou de l'autre.

Il reste à voir aussi comment utiliser concrètement cette information pour la désambiguïsation du verbe : a priori ce sont les compléments essentiels qui nous intéressent le plus, parce qu'ils définissent les constructions, mais rien ne dit que certains compléments circonstanciels, une fois identifiés comme tels, ne pourraient pas être utiles. Par exemple *jouer en public* influence fortement le sens de *jouer* (vers le sens *interpréter*). C'est donc un vaste chantier qui est ainsi ouvert.

9.2. Des classes de sélection distributionnelle[55]

Ce deuxième point d'amélioration consiste à finaliser notre modèle de désambiguïsation en tenant compte des caractéristiques sémantiques du co-texte lexical. Il s'agit d'associer une région de l'espace sémantique non plus à chaque co-texte comme nous l'avons proposé jusqu'à maintenant, mais à des classes de co-textes. Nous avons vu dans le chapitre 8, que notre modèle permettait de réduire fortement l'ambiguïté d'un verbe en combinant l'influence de co-textes lexicaux et syntaxiques. Cependant, on sait d'ores et déjà qu'il échoue sur des énoncés du type :

1) *Jouer du luth*
2) *Jouer à Wimbledon*

On se heurte ici à un double problème. Nous avons à faire à des compléments très peu représentés dans le corpus de référence. Or notre calcul repose sur l'utilisation des fréquences de cooccurrence de *luth* ou *Wimbledon* avec chacun des synonymes de *jouer*. Si ces fréquences sont trop faibles, le résultat du calcul est peu fiable. La première idée a été de remplacer *luth* par l'ensemble de ses synonymes, et de calculer leurs fréquences de cooccurrence avec chaque synonyme de *jouer*. Le problème est que les synonymes de *luth* sont trop peu nombreux et trop peu fréquents dans le corpus pour que le calcul soit efficace. Quand à *Wimbledon*, comme la plupart des noms propres, il ne possède aucun synonyme. Nous avons donc cherché à pallier ce manque d'informations quantitatives en fournissant à notre système des informations sémantiques sur les mots en question. Si nous pouvons associer à *luth* un ensemble de mots représentatifs des instruments de musique (*luth, guitare, piano, violon, etc*), nous retombons alors sur des énoncés interprétables par Visusyn et nous sortons de l'impasse. L'idée n'est certes pas nouvelle mais l'originalité de notre travail réside dans le fait que les classes que nous voulons construire vont dépendre du contexte dans

[55] Ce sous-chapitre reprend dans ses grandes lignes l'article de Jacquet et Venant (2005).

lequel le mot considéré est inséré. Par exemple pour le mot *luth*, dans le contexte « jouer du », la classe qu'on cherche à construire est celle des instruments de musique mais si on s'intéresse à l'énoncé *poser un luth*, la classe construite pour *luth* correspondra plutôt à une classe générale d'objets matériels. Le sens de *poser* dans *poser un luth* est en effet celui de *poser* dans *poser un objet* plutôt que celui de *poser* dans *poser ses congés* ou dans *poser une question*. Notre objectif, à terme, est de remplacer dans notre système de désambiguïsation les noms propres ou rares par leurs classes contextuelles et de retrouver ainsi le sens des verbes.

9.2.1 Méthode

La technique que nous utilisons s'inscrit dans le cadre de l'analyse distributionnelle « à la Harris ». Elle est exploitée dans la communauté du TAL pour la construction de bases de connaissances ou de ressources terminologiques à partir de textes (Frérot, 2003 ; Habert et Nazarenko, 1996 ; Fleury, 1998 ; Aussenac-Gilles *et al.*, 2000, Pantel et Lin, 2001, 2002). Notre méthode est entièrement automatique. Elle ne fait appel à aucune modélisation préalable de connaissances sémantiques sur le corpus et utilise des rapprochements de mots sur la base de contextes syntaxiques partagés. En tout cela, elle se rapproche des travaux de Greffenstette (1994). Les contextes nous sont fournis par l'analyseur Syntex (Bourigault et Fabre, 2000). Comme le précise D. Bourigault : « Là où Greffenstette se contente volontairement d'une analyse syntaxique relativement rudimentaire, réalisée par l'analyseur Sextant, nous avons fait le choix d'une analyse, certes encore partielle, mais plus large et plus précise, réalisée par Syntex. De ce fait, les procédures statistiques d'analyse distributionnelle de Greffenstette ne concernent que des mots simples, alors que nous pouvons prendre en compte des entités complexes (contextes ou termes) », cela nous permet de prendre en compte des distinctions plus fines, de créer des classes plus riches en information sémantique et donc plus efficaces dans leur apport à la désambiguïsation automatique. Notre travail est à rapprocher de celui de Habert *et al.* (2004) et Pantel et Lin (2001, 2002). Nous travaillons nous aussi à partir des rapports de dépendance syntaxique

élémentaire entre un contexte et les mots pleins qu'il régit ou qui le régissent et nous considérons les mots comme des points dans l'espace à n dimensions des contextes (que nous appelons l'espace distributionnel). Nous poursuivons cependant des objectifs différents. Nous ne cherchons pas à créer des classes de mots ayant le même sens mais des classes de mots dont le comportement sémantique influence de la même façon un contexte donné. Autrement dit si nous voulons trouver la classe de *luth (guitare, piano,…)* ce n'est pas pour caractériser le sens de *luth* mais pour désambiguïser *jouer* dans *jouer du luth*. Nous ne cherchons pas non plus à « faire parler le corpus dans sa globalité » comme le font Aussenac-Gilles *et al.* (2000). Les classes qu'ils construisent se constituent en naviguant autour d'éléments saillants ou prototypiques et leur permettent d'obtenir une image sémantique du corpus. Nous nous intéressons au contraire à des mots relativement peu fréquents, et qui ne représentent donc pas une ligne de force du corpus, pour rechercher dans leur classe sémantique des mots plus fréquents et dont l'apport à la désambiguïsation automatique sera plus pertinent. Certes les classes obtenues rendent compte de l'information sémantique présente dans le corpus mais de façon mouvante (Habert *et al.*, 1999). Chaque interrogation concerne un contexte et un mot différents et donne lieu à des regroupements différents au sein de l'espace distributionnel. Nos classes s'apparentent plutôt aux classes d'objet décrites par Gaston Gross (2004) : « tout changement de sens d'un prédicat est corrélé à un changement de son schéma d'arguments. Soit la phrase *Vous suivrez ce chemin*. Si on remplace l'objet *chemin* par des substantifs comme *route, rue, voie, sentier* le verbe *suivre* garde le même sens. On regroupera ces mots sous le terme générique de < voies >. Si en revanche, on remplace le mot *chemin* par *cours*, alors on a affaire à un autre emploi et le substantif *cours* peut être remplacé par *séminaire, stage, formation, cycle d'étude*, etc., qu'on rangera sous le classifieur < enseignement > ». Nous partageons avec Gross l'idée que « la mise au point du sens exige que l'on soit à même de préciser la nature sémantique des arguments que prend un emploi donné de prédicat » mais la différence entre les classes que nous cherchons et les classes d'objets de Gross, c'est que nous ne cherchons pas à établir des classes en langue. Nos classes dépendent du

contexte et surtout du corpus étudié. Gross cherche à créer des classes pouvant figurer dans un dictionnaire, c'est à dire calculées une fois pour toutes sur le lexique et indépendantes du corpus étudié. Nous proposons quelque chose de plus souple. Nos classes sont calculées en ligne pour désambiguïser un contexte dans un corpus donné. Elles ne sont valables que pour ce contexte même s'il peut y avoir des recouvrements. Elles ne sont pas nécessairement générales ni référentiables par un classifieur conceptuel comme < enseignement >. Elles caractérisent un comportement sémantique au sein d'un corpus donné plutôt qu'une notion et ne sont absolument pas hiérarchisables. L'intérêt de travailler à partir d'un contexte particulier est de limiter le nombre d'éléments à classer. Lorsqu'on étudie par exemple les compléments d'un verbe donné, on ne cherche pas à classer tous les noms de la langue française mais seulement les noms pertinents dans le contexte de ce verbe. Les classes sont obtenues plus facilement et sont plus significatives que des classes construites sur la globalité du lexique.

9.2.2 Constructions des CSD (Classes de sélection distributionnelle)

Données initiales

Nous travaillons sur le corpus LM10 qui, rappelons-le, regroupe tous les articles du journal *Le Monde* sur dix ans, soit 200 millions de mots. L'analyseur syntaxique Syntex (Bourigault et Fabre, 2000) dont nous avons déjà parlé (chapitre 8 § 8.1) nous fournit la liste des mots lemmatisés contenus dans le corpus avec leur fréquence ainsi que la liste des *triplets* {recteur ; relation ; régi} du corpus, avec leur fréquence. On a par exemple le triplet {compter(V) ; PREP_SUR ; ami(N)}[56] présent 13 fois dans le corpus et correspondant à des énoncés tels que *compter sur ses amis*. Il y a 20 millions de triplets différents (20 125 540 très exactement). Nous appellerons contexte lexico-syntaxique (C.L.S.) le couple formé par un des mots du triplet et la relation syntaxique.

[56] Pour les relations prépositionnelles, les deux triplets {compter(V) ; __ ; sur(Prep)} et {sur(Prep) ; __ ; ami(N)} sont fusionnés en un seul {compter(V) ; PREP_SUR ; ami(N)}.

Chacun des triplets va être séparé en un C.L.S. régi, un C.L.S. régissant, et deux mots. Le triplet {compter(V) ; PREP_SUR ; ami(N)} donnera ainsi deux C.L.S., « compter(V).PREP_SUR » présent 8 860 fois dans le corpus et « PREP_SUR.ami(N) » présent 88 fois et deux mots *compter(V)* et *ami(N)* de fréquences respectives 81 485 et 38 856. Nous obtenons ainsi une liste de mots (ou syntagmes) et une liste de contextes lexico-syntaxiques munis de leurs fréquences respectives. Ces listes constituent nos données de départ.

Données filtrées :

Les listes ainsi obtenues constituent une base de données colossale difficilement exploitable en l'état. Pour des raisons de taille et surtout de fiabilité, nous avons dû filtrer les informations qu'elle contient. Nous avons appliqué successivement les critères suivants : chaque mot et chaque C.L.S. doivent être présents au moins 100 fois dans le corpus et chaque triplet doit être présent au moins 10 fois dans le corpus.

Après filtrage notre ressource contient 31 417 mots et 61 202 contextes. A partir de ces données, nous construisons l'espace multidimensionnel engendré par les C.L.S.. C'est ce que nous appelons l'espace distributionnel associé au corpus. Chaque mot y est représenté par un point. La coordonnée d'un mot M sur l'axe engendré par un contexte C est la fréquence relative du triplet formé par M et C. Cet espace est muni de la distance du Chi2 : soit n le nombre de mots, p le nombre de contextes, M_i et $M_{k'}$ des mots de coordonnées (x_i^j) et (x_k^j) alors $d(M_i, M_k)^2 = \sum_{j=1}^{p} \frac{1}{x_\bullet^j} \left(\frac{x_i^j}{x_i^\bullet} - \frac{x_k^j}{x_k^\bullet} \right)^2$ où $x_\bullet^j = \sum_{i=1}^{n} x_i^j$ et $x_i^\bullet = \sum_{j=1}^{p} x_i^j$

Etude d'un mot dans un contexte lexico-syntaxique donné

Soient les énoncés *descendre le Mont-blanc* et *descendre la Seine*. Imaginons que nous cherchons à désambiguïser le verbe *descendre.* Une des forces de notre méthode est qu'elle permet d'étudier des cooccurrences non présentes dans le corpus.

Par exemple, on ne rencontre aucune occurrence de *Mont-blanc* ni *Seine* qui soit objet de *descendre*. Il est cependant possible d'étudier les mots *Mont-blanc* et *Seine* dans le C.L.S. « descendre.OBJ ». Nous allons d'abord chercher dans l'espace distributionnel tous les mots qui ont une coordonnée non nulle selon cette dimension, soit tous les mots filtrés employés dans ce contexte. On ajoute à la liste des mots obtenus les mots *Mont-blanc* et *Seine.* Notons que l'on fait une recherche toutes catégories confondues et que l'ensemble recherché peut contenir aussi bien des adjectifs, des noms communs, des noms propres ou même des entités plus complexes.

Si cet ensemble contient plus de 100 mots, on ne prend que les 100 mots les plus proches (au sens du Chi2) de *Mont-blanc* dans l'espace distributionnel. Notons MOTS l'ensemble formé. On va ensuite recenser tous les contextes pour lesquels au moins un des éléments de MOTS a une coordonnée non nulle. Notons CONT l'union de tous ces contextes. Dans le cas de *Mont-blanc*, MOTS contient 24 mots et CONT contient 5 762 contextes. Nous pouvons dans un premier temps visualiser l'ensemble MOTS grâce à une analyse factorielle des correspondances (AFC) qui nous fournit 10 axes de visualisation synthétisant le mieux l'information des 5 762 contextes de CONT.

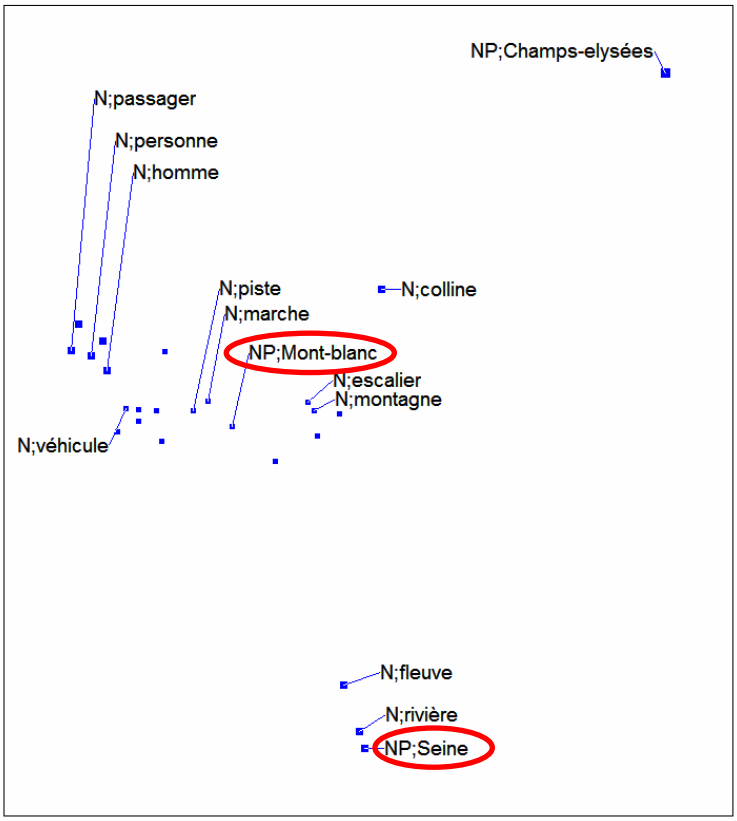


***Figure 9.4 : Distribution des mots compléments d'objet de* descendre *auxquels on a ajouté* Seine *et* Mont-blanc**

La figure 9.4[57] fait clairement apparaître trois axes sémantiques organisant les compléments d'objet du verbe *descendre* : les personnes, les monts ou surfaces inclinées, et les cours d'eau. Notons que *descendre un avion* se trouve entre *descendre une personne* et *descendre les marches*. Il est remarquable que *Seine* et *Mont-blanc* qui ne sont pas des compléments d'objet de *descendre* dans le corpus étudié trouvent automatiquement leur place le long de l'axe qui leur correspond le mieux. *Seine* est placé à coté de *rivière* et *fleuve*, alors que *Mont-blanc* est entouré de *piste*, *montagne* et *escalier*. La visualisation proposée correspond aux composantes 3 et 4 de l'AFC.

[57] Il est important de noter que malgré les apparences, cette visualisation n'a absolument rien à voir avec les espaces sémantiques que nous avons présentés dans les chapitres précédents. Nous avons maladroitement utilisé le même formalisme de visualisation mais il s'agit ici d'un espace distributionnel où les points sont des mots, et non un espace sémantique où les points sont des cliques.

Autrement dit, l'information est contenue dans l'ensemble des composantes de l'AFC, et obtenir une visualisation lisible nécessite de parcourir les différentes composantes. C'est pourquoi la construction des classes distributionnelles va tenir compte des dix premières dimensions de l'AFC. On pondère lors de la clusterisation chaque axe par le coefficient $1-((x-1)^2/100)$ (où x est le numéro de l'axe). Nous avons choisi la méthode de clusterisation k-means de Matlab (Seber, 1985). K-means emploie un algorithme itératif en deux phases dont le but est de minimiser la somme des distances entre points et centre de gravité sur le nombre k de clusters. Pour un mot M et un contexte lexico-syntaxique C donnés, notre modèle propose un ensemble de classes de sélection distributionnelle employées avec le contexte C. Nous avons la possibilité d'ordonner ces classes en fonction de leur proximité avec le mot M, la première classe sémantique étant celle qui contient le mot M. Il arrive parfois que M soit l'unique mot de la classe, dans ce cas nous considérons que la classe la plus proche de M est la seconde. Ainsi dans le contexte « descendre.OBJ » la classe la plus proche de *Seine* est «N;fleuve, N;rivière, NP;Seine » et la classe la plus proche de *Mont-blanc* est « N;montagne, N;piste ».

9.2.3 Evaluation des résultats

Nous proposons maintenant une première évaluation de notre système. Elle porte sur quatre contextes particulièrement ambigus en fonction des caractéristiques sémantiques de leurs arguments : « descendre.OBJ », « jouer.PREP_à », « regarder.OBJ », « décider.SUJ ». Pour chaque contexte, nous calculons la classe sémantique la plus proche de quinze mots vedettes différents. Voici la liste des 60 cooccurences étudiées. Nous avons marqué d'un [P] celles qui sont présentes dans le corpus :

« descendre.OBJ» : NP;Seine, NP;Rhône, NP;Gange, NP;Danube, NP;Mississippi, NP;Chirac, NP;Jospin, NP;Pdg, NP;Kennedy, NP;Mont Blanc, NP ;Everest, NP;Pyrénées, NP;Alpes, NP; Broadway

« jouer.PREP_à » NP;Monopoly [P] , N;tarot, N;domino [P] , N;lego, NP;Paris [P] , NP;Washington, P;Wimbledon [P] P;Lyon, NP;Broadway, NP;New York [P], NP;Londres [P], NP;Marseille [P], NP;Lille, NP;Parc des princes [P]

« décider.SUJ » :NP;Paris [P], NP;France [P], NP;Washington [P], NP;Wimbledon, NP;Londres [P], NP;Clinton [P], NP;président, NP;Jospin [P], NP; Kennedy, NP;Onu [P], NP;Cgt [P], NP;Otan [P], NP;Rpr [P], NP ;PS [P], NP;Vivendi, NP;Renault [P]

« regarder.OBJ » : NP;Chirac, NP;Picasso, NP;Seine, NP;Arte, NP; Kennedy, NP;Internet, NP;Alpes, NP;Jospin, NP;Paris, NP;Lyon, NP;Lelouch, NP;Kubrick, NP;Etats-Unis, NP;Tf[P], NP;Tintin, N;Videocassette

L'évaluation, inspirée des travaux de Lin et Pantel (2001), consiste à juger si la classe proposée par le modèle est acceptable ou non. Huit juges, dont nous nous sommes naturellement exclu, vont donner une note de un à quatre, de la manière suivante. La classe est très mauvaise : 1 ; La classe est assez mauvaise : 2 ; La classe est assez bonne : 3 ; La classe est très bonne : 4.

Contexte « descendre.OBJ »		
Mots vedettes	**Cluster le plus proche du mot vedette**	**Ev.**
NP;Seine	N;fleuve, N;rivière, N;Seine	4
NP;Gange	N;fleuve, N;rivière	3,9
NP;Chirac	N;homme, N;personne	3,8
NP;Pdg	N;homme, N;personne	3,1
NP;Mont Blanc	N;montagne, N;piste	3,5
NP ;Pyrénées	N;rivière, N;fleuve	1,6
Contexte « jouer.PREP_à »		
Mots vedettes	**Cluster le plus proche du mot vedette**	**Ev.**
NP;Monopoly	N;billard, N;bridge, N;cache-cache, N;domino, N;ping-pong, N;poker, N;pétanque, N;souris, N;yo-yo, NP;Monopoly	3,4
NP;Lego	N;billard, N;bridge, N;cache-cache, N;domino, NP;Lego, N;ping-pong, N;poker, N;pétanque, N;yo-yo, NP;Monopoly	3,5
NP;Paris	NP ;New York, NP;Avignon, NP;Londres, NP;Marseille, NP;Paris	3,5
NP;Washington	NP ;New York, NP;Londres, NP;Paris, NP;Washington	3,6
NP;Wimbledon	N;basket, N;basket-ball, N;football, N;loterie, N;rugby, N;tennis, S;jeu vidéo	1,9
NP;Broadway	N;base-ball, N;cricket, N;foot, N;golf, N;loto, N;volley-ball, NP;Broadway	1,4
Contexte « décider.SUJ»		
Mots vedettes	**Cluster le plus proche du mot vedette**	**Ev.**
NP;Washington	NP;Etats-Unis, NP;Washington	3.7
NP;Wimbledon	NP;Europe, NP;France, NP;Italie	2
NP;président	NP;Pdg, S;directeur général, S;président de le conseil, S;secrétaire de état	3.7
NP;Chirac	NP;Chirac, NP;Clinton, NP;Eltsine, NP;Jospin, NP;Mitterrand, S;chef de le état, S;premier ministre, S;président de le république	3.4
NP;Otan	N;armée, N;force, N;police	3.4
Contexte « regarder.OBJ »		
Mot vedettes	**Cluster le plus proche du mot vedette**	**Ev.**
NP;Chirac	N;enfant, N;femme, N;gens, N;homme, N;personne, N;public, NP;Chirac	3.1
NP;Seine	N;mer	2.8
NP;Arte	NP;Arte, NP;Tf, S;chaîne de télévision	3.8
NP;Alpes	N;montagne	3.8

Figure 9.5 : Extrait des notes d'évaluation

Nous avons détaillé quelques notes pour chaque contexte dans la figure 9.5. Dans le contexte « descendre.OBJ », on peut constater que les noms propres *Gange* et *Seine* ont dans leur première classe *fleuve* et *rivière*. Les noms propres *Chirac* et *PDG* ont dans leur première classe *homme* et *personne*. Dans le contexte « jouer.PREP_à », des noms propres comme *Lego* ou *Monopoly* sont rattachés à des classes de jeux alors que *Paris* et *Washington* sont rattachés à des noms de villes.

On peut noter des erreurs telles que « descendre.OBJ » avec *Pyrénées* qui donne des classes contenant *fleuve* et *rivière*. Ou encore « jouer.PREP_à » avec *Broadway*, qui donne une classe contenant des noms de sport.

	jouer.à	**descendre.Obj**	**regarder.OBJ**	**décider.Suj**	**TOTAL**
% 3 et 4	**81,63**	**80,61**	**65,18**	**83,93**	**77,62**
4	48,98	57,14	31,25	63,39	**50,00**
3	32,65	23,47	33,93	20,54	**27,62**
2	5,10	10,20	25,00	13,39	**13,81**
1	8,16	8,16	9,82	2,68	**7,14**
Non réponse	5,10	1,02	0,00	0,00	**1,43**
Moyenne des notes	3,30	3,20	2,90	3,40	**3,20**
Moyennes >3 (en %)	85,71	57,14	31,25	68,75	**60,00**

Figure 9.6 : Evaluation des classes obtenues pour 4 contextes, 60 cooccurrences, 8 juges.

La figure 9.6 propose une synthèse des résultats sur les quatre contextes. Les résultats de l'évaluation sont tout à fait satisfaisants. 77.62 % des notes sont supérieures ou égales à 3 et correspondent donc à des jugements de classe assez ou très bonnes. On peut noter que dans la moitié des cas la note mise est un 4. La moyenne des notes sur les 4 contextes est de 3,2 ce qui veut dire que les classes proposées sont globalement confirmées par les juges. Enfin 60% des classes ont sur l'ensemble des juges une note moyenne strictement supérieur à 3. On peut donc dire que 60% des classes proposées par notre système sont bonnes ou assez bonnes. 5 d'entre elles ont été jugées très bonnes par tous les juges. Ce sont les classes correspondant aux énoncés « descendre la Seine/ le Rhône/ le Danube » et « le PS/ la CGT décide ».

Il est par ailleurs intéressant d'étudier le comportement d'un même nom propre dans des contextes différents. Par exemple, *Wimbledon* peut être employé dans des énoncés tels que *revenir de Wimbledon*, *Wimbledon décide* ou *jouer à Wimbledon*. La figure 9.7 présente pour chacun des contextes correspondants, les deux classes les plus proches de *Wimbledon*. On observe que les différentes facettes de *Wimbledon* sont mises en évidence en fonction du contexte. Le contexte «décide.SUJ » met en valeur *Wimbledon* en tant que zone géographique de décision. Le contexte « jouer.PREP_à» insiste sur la singularité de *Wimbledon* en tant que compétition de tennis. Les classes obtenues pour le contexte « revenir.PREP_de » peuvent être interprétées comme des lieux d'activité.

	Wimbledon dans le contexte « Wimbledon décide»	**Wimbledon** dans le contexte « jouer à Wimbledon »	**Wimbledon** dans le contexte « revenir de Wimbledon »
1er cluster	NP;Europe, NP;France, NP;Italie	N;basket, N;basket-ball, N;football, N;loterie, N;rugby, N;tennis, S;jeu vidéo	NP;Allemagne, NP;Etats-unis
2ème cluster	N;monde, N;pays, N;région, N;ville	N;base-ball, N;cricket, N;foot, N;golf, N;loto, N;volley-ball	N;guerre, N;mission, N;travail

Figure 9.7 : Différentes classes de* Wimbledon *en fonction du contexte

9.2.4 Discussion

Nous avons présenté ici une méthode de construction de classes de sélection distributionnelle en contexte. Cette méthode, au vu des résultats préliminaires présentés ici, nous semble très prometteuse. Des expérimentations plus poussées sont cependant encore nécessaires pour finaliser notamment la méthode de clusterisation (détermination du nombre de clusters optimal, pondération des axes de l'AFC) ou le mode de filtrage du corpus. Il nous faudrait valider sur un plus grand nombre de contextes, sur différentes catégories de mots et en faisant appel à un plus grand nombre de juges. Nous devons aussi étudier la variation des classes obtenues lorsqu'on change le corpus de travail. Par exemple il est clair que la classe « *NP;Chirac, NP;Clinton, NP;Eltsine, NP;Jospin, NP;Mitterrand* » ne se retrouvera pas dans le corpus Frantext. Ni même

dans un corpus composé des trois dernières années du journal *Le Monde*. En revanche, il est probable que certaines classes moins connotées telles que « *N;enfant, N;femme, N;gens, N;homme, N;personne* » traversent les frontières inter corpus (Habert *et al.*, 1999 ; Manguin *et al.*, 2005).

Le type d'évaluation que nous avons mis en place n'est qu'une étape. Le but à terme est d'utiliser les classes obtenues, dans notre système de désambiguïsation. Il nous faut donc mettre au point le module de Visusyn correspondant. Nous devrons déterminer quelle est la façon la plus pertinente de prendre en compte les classes de sélection distributionnelle dans Visusyn.

10. Conclusion et perspectives

L'ensemble des études que nous avons développées peut paraître hétérogènes. Cette partie a pour but de cimenter cet ensemble dans un système global de désambiguïsation, et plus précisément de calcul du sens. Comme nous l'avions annoncé au début de cette thèse, l'objectif visé à long terme est le calcul du sens d'un énoncé. Nous avons assimilé cette tâche au calcul, pour chaque unité lexicale de l'énoncé, des synonymes équivalents ou proches, et à la production d'une série de paraphrases pour l'énoncé. Pour le moment nous avons présenté essentiellement une méthode de calcul des synonymes proches d'une unité lexicale dans un énoncé. Nous allons voir maintenant que nous avons aussi ouvert la voie à la construction de paraphrases. Nous commencerons par présenter quelques points de méthode que l'on pourrait améliorer. Nous présenterons ensuite quelques cas d'applications possibles et notamment une application pour le calcul de paraphrases d'une construction verbale. Nous verrons que les classes de sélection distributionnelles que nous avons présentées nous ouvrent des perspectives qui dépassent la tâche de la désambiguïsation verbale. Enfin nous conclurons par une perspective plus globale.

10.1. Améliorations méthodologiques

La tâche de calcul des synonymes les plus proches d'un verbe en fonction de l'influence de son co-texte, qu'il soit lexical ou syntaxique, est opérationnelle. Nous avons aussi montré qu'il était possible de combiner les influences lexicales et syntaxiques. Nous proposons ici quelques améliorations que l'on pourrait apporter au modèle. Le premier point développé revient sur quelques questions de méthode,

évoquées durant cette étude. Les deux points suivants proposent des améliorations plus importantes.

10.1.1 Points algorithmiques

Commençons par le calcul de l'information mutuelle. Nous nous sommes inspirés de ce calcul notamment pour déterminer le degré d'affinité entre le mot étudié et une unité co-textuelle. Nous avons déjà évoqué le fait que ce calcul était très sensible aux faibles fréquences. La solution pourrait être de supprimer les fréquences trop faibles avec un seuil à déterminer. Une autre solution serait de pondérer notre calcul basé sur l'IM. C'est-à-dire que, en plus du degré d'affinité que nous avons proposé, chaque clique aura aussi un degré de fiabilité en fonction des fréquences qui entrent en jeu dans le calcul du degré d'affinité de cette clique. Cette deuxième solution nous semble plus intéressante car elle permettrait de mettre en évidence les problèmes de fiabilité de calcul, plutôt que de les dissimuler. Le t-score semble être un indice adapté pour mesurer un tel degré de fiabilité[58].

Revenons maintenant sur le calcul des fonctions potentielles. Celles-ci sont projetées sur un espace à deux dimensions. Or nous avons vu qu'initialement, l'espace sémantique d'un mot contient autant de dimensions que ce mot possède de synonymes. Nous avons proposé de faire ce calcul sur un nombre plus important de dimensions. La première idée serait de faire le calcul sur les *n* dimensions initiales. Cette idée pourrait sembler mieux correspondre à la réalité, or ce n'est certainement pas la meilleure solution. En effet, l'intérêt de l'AFC est justement de faire émerger les tendances fortes de l'espace sémantique. Une solution plus envisageable serait de calculer la fonction

---

[58] Calcul du *t-score* développé dans Fung et Church (1994) :

$$t \approx \frac{p_{12} - p_1 p_2}{\sqrt{\frac{p_{12}}{n}}}$$ où $p_1 = \frac{freq(u_1)}{n}$, $p_2 = \frac{freq(u_2)}{n}$, $p_{12} = \frac{freq(u_{12})}{n}$ et n le nombre de phrases total.

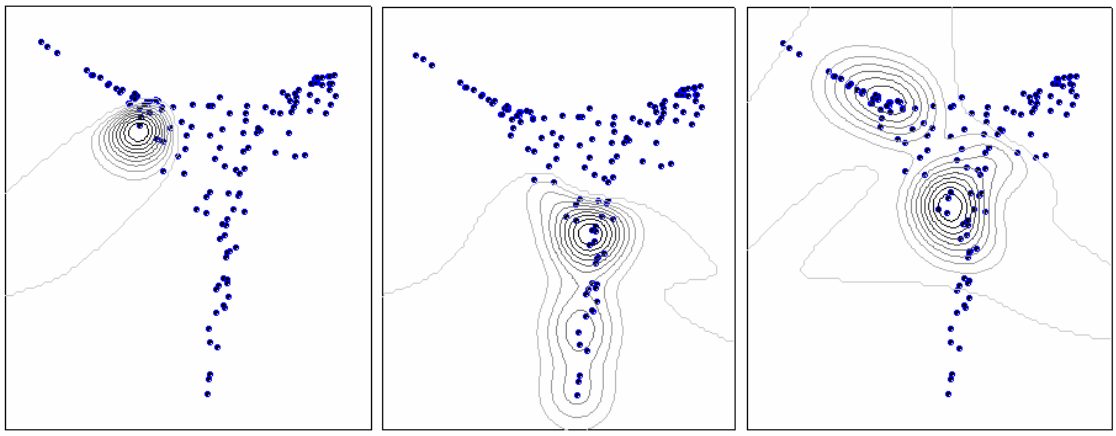
potentielle sur les trois ou quatre premières dimensions de l'AFC. Une autre solution serait de faire le calcul sur les deux dimensions les mieux adaptées à la région de sens à activer. Cela reviendrait à avoir un espace sémantique personnalisé pour chaque requête (ce qui n'est pas sans poser de problèmes de visualisation). Si l'on reprend le cas du verbe *monter*, le sens *bâtir/construire/monter une maison* se trouve au milieu de l'espace sémantique que nous avons présenté (§ 4.3.1). Ce sens se trouve dans la même région que tous les sens qui ne sont pas représentés dans les trois axes de sens dominants *grimper, escalader ; manigancer, préparer ; augmenter, rehausser*. Lorsque l'on étudie le sens de *monter* dans *monter une maison*, l'idée serait alors d'augmenter le poids des verbes ayant des affinités avec *maison* (*bâtir* et *construire*) pour que les deux premières dimensions de l'AFC fassent émerger des axes de sens contenant ces verbes. D'une certaine manière cela correspond à passer d'une représentation du sens en langue générale à une représentation dépendante du contexte d'emploi.

10.1.2 Les corpus

Les différentes études que nous avons présentées sont toutes basées sur corpus. Nous avons utilisé trois corpus assez différents, Frantext, LM3 et LM10. Les deux derniers sont proches par leur contenu puisqu'ils proviennent du même journal (Le Monde) mais ils diffèrent dans leur format : nous avions une analyse syntaxique complète pour le LM3 mais uniquement des triplets de relations syntaxiques pour le LM10 (triplet : recteur ; relation ; régi). Nous pourrions considérer le Web, que nous avons exploité dans le paragraphe 9.1 comme un quatrième corpus. Pour calculer l'influence du co-texte syntaxique, nous avons utilisé le seul corpus qui nous permettait ce type de calcul, c'est-à-dire le LM3. Nous avons vu dans le chapitre 6 que notre calcul de degré d'affinité étant très sensible aux faibles fréquences. On peut alors se demander s'il ne serait pas intéressant de faire nos calculs sur un corpus plus important. Nous avons conscience que le problème ne serait que déplacé, c'est-à-dire que la répartition des fréquences suivrait la même loi de Zipf, néanmoins le nombre de calculs ne faisant intervenir que des fréquences « fiables » augmenterait. La question est de savoir

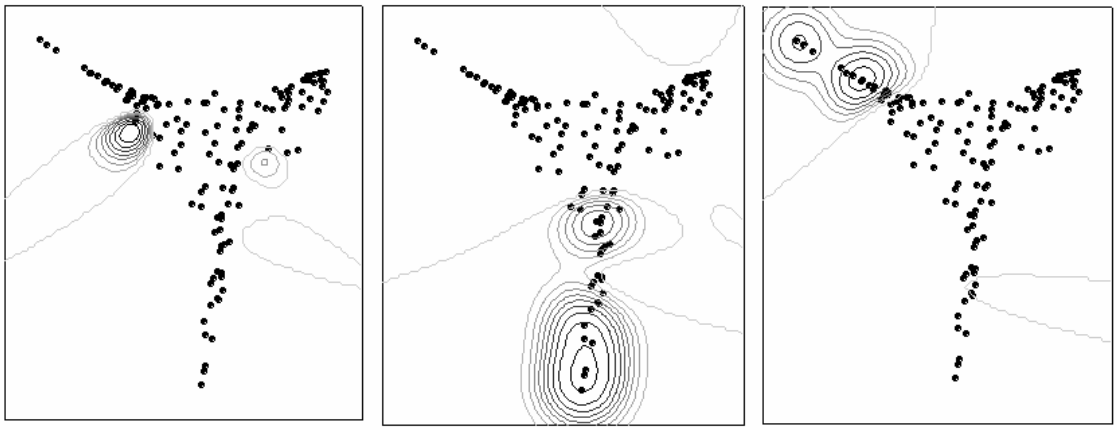
jusqu'où on peut se permettre d'augmenter la taille d'un corpus ? Et surtout avec quels textes ?[59]. Par exemple, pourrait-on se permettre de réunir Frantext et Le Monde ? A première vue, nous aurions tendance à répondre par la négative étant donné les nombreux points qui opposent ces deux corpus (styles, thèmes, période, etc.). Cependant il semble intéressant de montrer que les fonctions potentielles calculées pour l'influence d'un co-texte lexical sur le sens d'un verbe soient assez proches selon que l'on utilise Frantext ou le corpus LM3. On peut notamment observer figure 10.1 et 10.2, la comparaison entre les deux corpus pour les fonctions potentielles des co-textes *diamant*, *escalier* et *projet* sur l'espace sémantique de *monter*. L'étude faite par Manguin (2005) sur le calcul de distances entre cooccurrences semble aussi montrer des similitudes entre les résultats obtenus à partir d'un corpus journalistique (regroupement d'articles des journaux *Le Monde, Libération, Le Point*) et d'un corpus littéraire (Frantext).

[59] Voir Habert *et al.* (1997 : chap. 7) pour un point développé sur ces questions.



***Co-texte* diamant** ***Co-texte* escalier** ***Co-texte* projet**

Figure 10.1 : Fonctions potentielles obtenues à partir du corpus Frantext



***Co-texte* diamant** ***Co-texte* escalier** ***Co-texte* projet**

Figure 10.2 : Fonctions potentielles obtenues à partir du corpus LM3

10.1.3 Les constructions

Revenons maintenant sur l'étude des constructions syntaxiques. Nous avons soulevé les problèmes liés au fait que notre analyseur syntaxique ne distingue pas les compléments essentiels des circonstanciels. Nous avons proposé une méthode pour attribuer une note aux compléments étudiés en fonction de leur degré d'essentialité, ou plutôt de leur importance pour la désambiguïsation du verbe. Cette méthode est encore à développer et à évaluer convenablement. En admettant que les résultats obtenus soient satisfaisants, ce que nous espérons, reste encore à déterminer comment intégrer cette

méthode dans notre modèle de désambiguïsation. On pourrait imaginer une pondération des compléments en fonction de leur degré d'influence sur le verbe, mais la méthode reste à mettre au point.

Cette pondération est importante pour le développement de notre modèle. D'une part elle permettrait de réduire les erreurs de fonction potentielle telles celles que nous avons observées pour *compter sur SN* qui active par erreur la région *énumérer, calculer* et *jouer de SN* qui active des régions correspondant à des énoncés tels que *jouer de façon remarquable.* Cette pondération devient indispensable si nous voulons pouvoir calculer l'influence de constructions syntaxiques plus complexes que celles que nous avons étudiées ici. Par exemple, la construction ditransitive *V SN1 à SN2* en français nécessite de savoir si les deux syntagmes *SN1* et *à SN2* sont bien des compléments d'objet du verbe *V*.

10.1.4 Combinaison lexique - syntaxe

Nous avons présenté une manière de faire interagir l'influence du lexique et de la syntaxe en additionnant les fonctions potentielles des deux types de co-textes. Les résultats semblent correspondre à nos attentes, néanmoins cette opération de combinaison mérite une étude plus approfondie. La question principale est de savoir s'il vaut mieux calculer l'influence d'un co-texte lexical en contraignant ce co-texte par la relation syntaxique qu'il entretient avec le mot à désambiguïser, c'est-à-dire en faisant de la sémantaxe (François, 2003, et aussi chapitre 2 §2.5) ou bien séparer les calculs de l'influence du lexique et de la syntaxe. L'avantage de la sémantaxe est de traiter une construction verbale dans son ensemble et donc de pouvoir déterminer directement l'influence de cette construction sur le sens du verbe. L'inconvénient apparaît en terme de calcul. En effet, plus la construction verbale sera complexe (sujet + complément avec les éléments lexicaux correspondants. Par ex. : *Notre équipe a monté un projet*), moins on aura de chance de la retrouver en tant que construction d'un des synonymes du verbe étudié. De plus, la sémantaxe ne permet pas de considérer les constructions syntaxiques comme des unités linguistiques à part entière, à la Goldberg. Notre choix s'était donc

initialement orienté vers une séparation du lexique et de la syntaxe. L'apport des classes de sélection distributionnelles (CSD) a remis en question ce choix. En effet, le fait de remplacer une unité lexicale par sa CSD permet de baser les calculs non sur la fréquence du co-texte lexical mais sur la somme des fréquences des unités appartenant à sa CSD. On peut ainsi repousser l'argument des problèmes de fréquence à des constructions encore plus complexes (par ex. : *les enfants ont donné du pain aux chevaux).* Et même sans tenir compte des CSD, certaines collocations ne méritent pas d'être décomposées. Par exemple, il serait contre-productif de décomposer *monter les escaliers* en l'influence de la construction transitive sur le sens du verbe *monter*, combinée à l'influence du co-texte lexical *escalier*.

En revanche, la sémantaxe ne permet pas de rendre compte directement de l'influence d'un co-texte syntaxique sur le sens du verbe. De plus, on peut se demander si remplacer chaque unité lexicale d'un énoncé par sa CSD, comme nous venons de le proposer, n'apporte pas plus de bruit que de précision.

Le choix semble difficile et la meilleure solution est probablement de combiner les deux approches en fonction de l'énoncé à traiter.

D'autres phénomènes syntaxiques doivent encore être traités. Notamment les cas de non compositionnalité[60]. Par exemple, *jouer un rôle* prend le sens de *interpréter un rôle* dans l'énoncé *jouer un rôle dans cette pièce de théâtre*. Mais il prend le sens de *influer*, et non ?*influer un rôle*, dans l'énoncé *jouer un rôle dans cette société*. Autrement dit, si l'on veut rendre compte du fait que *influer* est synonyme de *jouer un rôle* (et non simplement de *jouer* du moins pas dans cet énoncé) il faut traiter *jouer un rôle* comme il se doit, c'est-à-dire comme un syntagme verbal. Actuellement, nous sommes capable de calculer automatiquement que les synonymes de *jouer* dans *jouer un rôle* sont *interpréter* et *influer*, mais nous ne pouvons pas mettre en évidence le fait que c'est le syntagme verbal *jouer un rôle* qui est synonyme de *influer* et non

[60] Voir Nazarenko (1998) pour différentes études sur ce thème, sous l'angle du TALN.

simplement *jouer*. Or il existe de nombreux syntagmes verbaux de ce type. Néanmoins, l'analyseur syntaxique que nous utilisons, Syntex, a l'avantage de mettre en évidence des syntagmes de manière assez large. On retrouve des syntagmes tels que « *bilan de l'année* », « *intervention de l'Otan* ». Cette extraction de syntagme est automatique mais limitée aux syntagmes nominaux. On pourrait imaginer appliquer la même méthode à des syntagmes verbaux afin de pouvoir traiter ce type de cas.

10.2. Applications

A partir du moment où nous développons un modèle de désambiguïsation automatique, les applications en TAL sont importantes. Améliorer les moteurs de recherche en ne retournant que les résultats pertinents par rapport au sens de la requête, optimiser l'indexation de document en n'indexant plus à l'aide de mots ou groupes de mots mais à l'aide de groupes de sens, etc. Nous commencerons par présenter ce que nous avons appelé le calcul de paraphrase de constructions verbales. Nous développerons ensuite quelques applications.

10.2.1 Paraphrases de constructions verbales

Une application directe serait non plus de proposer une liste de synonymes pour chaque unité lexicale de l'énoncé mais une liste de synonymes dans une construction donnée. L'idée sous jacente est de pouvoir dire par exemple que ce n'est pas *pratiquer* qui est synonyme de *jouer* dans *jouer de la guitare*, mais c'est « pratiquer.OBJ » qui est synonyme de « jouer.PREP_DE ». Cela correspond à l'objectif que nous nous étions fixé au début de l'étude.

Nous proposons d'utiliser l'espace distributionnel que nous avons présenté dans le paragraphe 9.2, dans lequel, rappelons le, la coordonnée d'un mot M sur l'axe engendré par un contexte C est la fréquence relative du triplet formé par M et C. Cet espace est muni de la distance du χ^2. Mais nous n'allons plus nous intéresser à la distribution des mots en fonction des contextes avec lesquels ils cooccurrent comme

c'était le cas pour la construction des CSD, mais plutôt à la distribution des contextes en fonction des mots. En clair, nous faisons le même calcul sur l'espace dual de l'espace distributionnel initial. En revanche, nous n'utiliserons pas la méthode de clusterisation puisque la liste des contextes classés par ordre de distance avec le contexte étudié est suffisante ici. Nous extrayons ensuite de cette liste uniquement les contextes contenant un des synonymes du mot étudié.

Reprenons le cas de *jouer de la guitare*. Cet énoncé se décompose en un contexte « jouer. PREP_DE » et un mot *guitare*. La première colonne de la figure 10.3 correspond à la liste des contextes les plus proches de « jouer. PREP_DE » lorsqu'il est employé avec *guitare*. La troisième colonne correspond uniquement aux contextes contenant un des synonymes de *jouer*. On retrouve ainsi deux contextes correspondant à nos attentes : « pratiquer.OBJ » et « gratter.OBJ », c'est-à-dire *pratiquer la guitare, gratter la guitare*. On peut aussi noter que les contextes de la deuxième colonne, qui correspondent aux contextes contenant un verbe sont déjà pertinents (*accompagner à la guitare, utiliser la guitare, apprendre la guitare,* etc.). Il semble que l'on puisse, avec ce calcul proposer un type de relation de synonymie plus précis. On pourrait même se demander si le calcul de désambiguïsation par l'influence du co-texte est encore nécessaire. On retrouve en effet les synonymes *pratiquer* et *gratter* sans avoir à calculer de fonctions potentielles. Nous réservons notre réponse sur ce point. En effet, un synonyme de *jouer* ne correspondant pas au sens attendu aurait très bien pu faire partie de la liste pour d'autres raisons. Il est probable qu'il ne faille extraire de la liste des contextes que ceux contenant un synonyme retenu comme pertinent par notre modèle de désambiguïsation.

1.V;jouer.de	0
2.A;acoustique.EPI	0.21
1.N;son.EPI	0.40
1.N;solo.EPI	0.84
1.V;accompagner.à	0.88
2.A;électrique.EPI	0.97
1.V;pratiquer.OBJ	1.01
2.A;classique.EPI	1.09
1.N;riff.EPI	1.12
2.A;hawaïen.EPI	1.12
1.V;utiliser.OBJ	1.25
1.V;apprendre.OBJ	1.51
1.V;accompagner.SUJ	1.62
1.N;partie.EPI	1.65
2.A;espagnol.EPI	1.66
1.V;entendre.OBJ	1.67
2.A;gros.EPI	1.81
2.A;portugais.EPI	1.90
1.V;gratter.OBJ	1.94
1.N;accord.EPI	10.78
2.A;bas.EPI	14.19
1.V;accompagner.de	14.25
2.N;bandoulière.en	18.71
1.N;jeu.EPI	2.22
1.N;joueur.EPI	2.63
2.A;saturer.EPI	2.76
1.N;corde.EPI	21.32
2.A;rythmique.EPI	29.27
1.N;fond.EPI	3.41
1.N;joue.EPI	5.80

1.V;jouer.de
1.V;accompagner.à
1.V;pratiquer.OBJ
1.V;utiliser.OBJ
1.V;apprendre.OBJ
1.V;accompagner.SUJ
1.V;entendre.OBJ
1.V;gratter.OBJ
1.V;accompagner.de

1.V;jouer.de
1.V;pratiquer.OBJ
1.V;gratter.OBJ

Figure 10.3 : Processus de sélection de paraphrases d'une construction verbale

10.2.2 Applications directes

Les intérêts de cette synonymie contrainte par la construction sont multiples. Regardons l'intérêt pour un **moteur de recherche**. L'inconvénient classique des moteurs de recherche est de ne retourner que les documents contenant au moins un des mots de la requête. Cela implique un bruit important : tous les documents contenant un mot de la requête mais pas dans le sens voulu ; et un silence important : tous les documents ne contenant pas l'un des mots de la requête mais qui sont pertinents pour la requête. A l'inverse, retourner uniquement les documents contenant un des synonymes des mots de la requête, dans le sens de la requête fournira un résultat très précis mais qui nécessite un temps de calcul irréaliste. En effet, cela nécessiterait de désambiguïser

chaque occurrence (dans la base de documents) de chaque synonyme de chaque mot de la requête. Ce que nous proposons semble un bon compromis puisque cela ne nécessite qu'une désambiguïsation des mots de la requête. Le résultat de cette désambiguïsation serait, grâce à la méthode que nous venons de présenter, une liste de synonymes dans une construction donnée. Cela permettrait de retourner uniquement les documents contenant l'un de ces synonymes, dans une construction précise.

Autrement dit, si la requête est « *jouer de la guitare* », on ne recherchera que les documents contenant *jouer de la guitare* ou *pratiquer la guitare* ou *gratter la guitare*. Cette méthode implique une contrainte : que l'ensemble des documents de la base soient analysés syntaxiquement. De ce fait, cette application semble difficile pour le Web, mais tout à fait réalisable pour des bases de documents même importantes (par exemple, le moteur de recherche développé par Synomia fonctionne à partir de documents analysés syntaxiquement à l'aide de Syntex et est opérationnel notamment sur l'ensemble des archives du journal *Libération*).

Application pour la **traduction automatique** : prenons le cas d'un mot ambigu à traduire, c'est-à-dire qui peut être traduit de différentes manières. L'idée serait dans un premier temps de désambiguïser le mot à traduire. On obtiendrait ainsi un petit nombre de synonymes, dans une construction donnée. Ensuite, utiliser un corpus aligné pour retrouver comment est traduit le mot ambigu dans sa construction ainsi que ses synonymes contraints par leur construction. L'idée reste la même c'est-à-dire être le plus précis possible dans la relation de synonymie pour que l'apport des synonymes soit supérieur au bruit qu'ils pourraient apporter.

Enfin, ces synonymes contraints pourraient tout simplement servir à construire automatiquement des **paraphrases basiques** pour un utilisateur qui ne comprend pas un énoncé :

« - Gratter la guitare.

- Quoi ?

- Jouer de la guitare, pratiquer la guitare.

- Ah ! D'accord… »

10.2.3 Retour à l'analyse syntaxique

Le module de distinction entre compléments essentiels et circonstanciels pourrait être directement ajouté aux règles d'un analyseur syntaxique. Comme nous l'avons dit dans le paragraphe 9.1, des tests similaires sont déjà exploités dans Syntex pour les ambiguïtés de rattachement prépositionnel, mais ce n'est pas de cela dont nous voulons parler ici. L'idée est plutôt de faire avancer l'analyse syntaxique de chaque énoncé en même temps que l'analyse sémantique. C'est à dire faire interagir les deux processus avec l'idée que les résultats de l'un aident aux choix de l'autre. Par exemple avec un énoncé tel que *monter ce matin à Paris*. On pourrait imaginer commencer par une analyse syntaxique basique, consistant simplement à repérer les différents syntagmes de l'énoncé. Le processus de désambiguïsation permet alors de montrer que *monter* prend le sens de *aller*, *se rendre* et que c'est le syntagme *à Paris* qui a permis de réduire l'ambiguïté de *monter*. En revanche le syntagme *ce matin* n'a rien apporté. On peut alors retourner vers l'analyseur syntaxique et favoriser le rattachement de *à Paris* au verbe *monter*.

De la même manière, on pourrait envisager de remonter jusqu'à l'étiqueteur grammatical. Les étiqueteurs tels que Tree Tagger ont un taux d'erreur très faible, mais la moindre erreur va se répercuter sur l'analyse syntaxique, puis sur les CSD que nous calculons, ce qui est plus problématique. Par exemple *revenu* est marqué comme une flexion du verbe *revenir* par tree Tagger dans *revenu de mille euros*. L'idée serait alors

de faire un premier passage de l'étiquetage morphosyntaxique et de l'analyse syntaxique en favorisant le silence. Puis faire un premier calcul de classes de sélection distributionnelle afin de désambiguïser les mots non encore étiquetés. Ainsi, le contexte ambigu « revenu.PREP_DE » sera rapproché de verbes de type « venir.PREP_DE », « aller.PREP_à » lorsqu'il est avec le mot *Paris*, alors qu'il sera rapproché de contextes tels que « salaire.PREP_DE » « montant.PREP_DE » lorsqu'il est avec le mot *euros*. On pourrait alors favoriser l'étiquetage en tant que verbe dans le premier cas, et en tant que nom dans le second.

10.3. Au delà de la désambiguïsation des verbes

Initialement, nous avons construit les CSD dans l'objectif d'améliorer notre modèle de désambiguïsation des verbes. Nous nous sommes rendu compte en les manipulant qu'elles pouvaient être utiles pour d'autres tâches d'analyse sémantique d'un texte. Notamment la résolution des anaphores réclame, dans certains cas, de connaître la classe sémantique à laquelle peut appartenir un argument d'un verbe donné. Par exemple, dans les phrases suivantes : *La voiture a renversé Marie. Elle a été blessée assez gravement.*, il faut absolument savoir que le sujet de *être blessé* doit être un humain ou un animal pour trouver que l'antécédent de *Elle* est *Marie* et non pas *la voiture.* Notre méthode devrait pouvoir être appliquée à profit à ce genre de problèmes en fournissant les classes de sélection distributionnelles pertinentes chaque fois que l'on en a besoin. Nous proposons ici de détailler deux autres cas d'utilisation.

10.3.1 CSD et entités nommées

Nous avons vu dans le paragraphe 9.1 que nos CSD permettaient de déterminer l'influence d'une entité nommée telle que *Seine* ou *Mont-blanc* sur le sens du verbe *descendre*. Mais les classes calculées ont un intérêt en soi. *Seine* et *Mont-blanc* soit des entités courantes et il est facile de les répertorier dans des classes de type *cours d'eau* et *montagne*. L'intérêt des CSD est qu'elles permettent de replacer une nouvelle entité nommée dans l'une de ces classes. Un nom de montagne qui n'aurait pas été répertorié

dans la liste des montagnes aura le même comportement distributionnel que les montagnes qui ont déjà été répertoriées. En cela, nos CSD ne sont pas plus efficaces que des classes distributionnelles classiques. En revanche, les CSD ont un intérêt important pour traiter les entités nommées ambiguës. Prenons le cas du mot *Orange :* ci-dessous les CSD calculées en fonction de différents contextes.

	orange dans le contexte « orange rachète»	**orange** dans le contexte « manger des oranges»	**orange** dans le contexte « acheter orange »
1er cluster	NP;Elf, NP;Renault, NP;Thomson	N;fruit	N;matière, N;produit
2ème cluster	N;holding	N;pomme	N;espace, N;modèle, N;programme
3ème cluster	N;banque, N;compagnie	N;riz	N;action, N;activité

On peut constater que *orange* dans le contexte « racheter.SUJ » est considéré comme une société (NP;Elf, NP;Renault, NP;Thomson, N;holding, N;banque, N;compagnie), alors que dans le contexte « manger.OBJ», *orange* est rapproché d'aliments (N;fruit, N;pomme, N;riz). Le cas du contexte « acheter.OBJ » est intéressant car ambigu. Il peut correspondre à des énoncés tels que *(1) acheter des oranges chez le marchand* ou au contraire dans un contexte boursier *(2) acheter Orange était une bonne affaire en 2001*. Or, dans ce contexte, les trois clusters les plus proches de *orange* ne véhiculent pas les mêmes notions. Si le deuxième cluster reste assez flou (N;espace, N;modèle, N;programme), on pourrait facilement dire que le premier cluster (N;matière, N;produit) correspond à l'emploi de *orange* dans l'énoncé (1), et le troisième cluster (N;action, N;activité) à l'emploi de *orange* dans l'énoncé (2).

On pourrait alors envisager d'utiliser ces classes pour replacer une entité nommée ambiguë dans la classe correcte en fonction de son contexte d'emploi. Ceci en respectant le fait que certains emplois, tels que « acheter.OBJ.orange » sont ambigus et doivent être classés comme étant ambigus.

10.3.2 CSD et désambiguïsation nominale

Nous avons montré l'intérêt des CSD pour traiter des co-textes rares (*jouer du luth*). On pourrait sans difficulté envisager de généraliser à l'ensemble des co-textes quelle que soit leur fréquence. L'intérêt étant de pouvoir traiter, de la même manière que pour les entités nommées, les noms ambigus. Nous n'aurons pas la même classe pour *bureau* dans *travailler sur le bureau* (*bureau, table, chaise*) que dans *entrer dans le bureau* (*bureau, cuisine, salon*). Cette distinction permettra de faciliter la désambiguïsation des verbes, respectivement *travailler* et *entrer*. On pourrait même envisager d'utiliser les CSD que nous venons de citer pour la désambiguïsation des noms en tant que tel. Cela reviendrait à dire que *bureau* dans l'énoncé *travailler sur le bureau* prend le sens de la classe (*bureau, table, chaise*) que l'on pourrait nommer '*meuble*'. Alors que dans *entrer dans le bureau*, *bureau* prend le sens de la classe (*bureau, cuisine, salon*) que l'on pourrait nommer '*pièce'*.

Mais c'est là que le bât blesse puisque la limite de nos CSD, et d'une manière générale de toutes les classes construites automatiquement à partir de corpus, est qu'elles peuvent proposer une classe de meubles telle que *bureau, table, chaise*, mais sans pouvoir nommer cette classe *meuble.* Cette tâche semble difficilement réalisable de manière automatique (Nazarenko *et al.*, 2001). Cependant, plusieurs travaux proposent des méthodes endogènes que l'on pourrait voir comme des aides à la dénomination de classes. Le principe consiste à rechercher en corpus des patterns particuliers pouvant correspondre à des hyperonymes : par exemple *X est un Y* [61]. Pantel et Ravichandran

[61] Liste des patterns syntaxiques utilisés par Pantel et Ravichandran :

(2004) proposent de rendre plus robuste cette méthode automatique avec le principe suivant :

On assigne l'hyperonyme *meuble* à une instance telle que *guéridon*, non parce que *guéridon* cooccurre nécessairement dans un pattern avec *meuble*, mais parce qu'il appartient à une classe d'instances dont les éléments les plus représentatifs (*bureau, table, chaise*) cooccurrent dans un pattern avec *meuble*. Pantel et Ravichandran obtiennent un taux de réussite très encourageant (68% de relations hyponymiques correctes pour les noms, 81,5% pour les noms propres) avec une évaluation à très grande échelle. Nous pourrions assez facilement appliquer cette méthode à nos CSD et ainsi proposer un hyperonyme calculé automatiquement pour chaque CSD, ce qui, nous l'avons vu serait très utile dans différentes applications.

10.4. Vers un modèle global de la désambiguïsation d'un énoncé

Nous avons présenté dans cette étude un modèle de désambiguïsation pour les verbes. Ce modèle permet de prendre en compte l'influence du lexique et de la syntaxe sur le sens du verbe. Nous avons même montré qu'il était possible de désambiguïser un verbe non par une liste de synonymes mais par une liste de synonymes dans une construction donnée. Parallèlement, nous venons de montrer que l'on pourrait envisager de désambiguïser les noms, non par une liste de synonymes mais plutôt par une CSD

- *Apposition (N:appo:N)*
  e.g. ... **Oracle**, a **company** known for its progressive employment policies, ...
- *Nominal subject (-N:subj:N)*
  e.g. ... **Apple** was a hot young **company**, with Steve Jobs in charge.
- *Such as (-N:such as:N)*
  e.g. ... **companies** such as **IBM** must be weary ...
- *Like (-N:like:N)*
  e.g. ... **companies** like **Sun Microsystems** do no shy away from such challenges, ...

dépendant du contexte, et nommée automatiquement. Cela reviendrait à désambiguïser les noms à l'aide d'une liste de co-hyponymes chapeautée par un hyperonyme. Nous approchons donc progressivement notre objectif à long terme qui est de calculer le sens d'un énoncé.

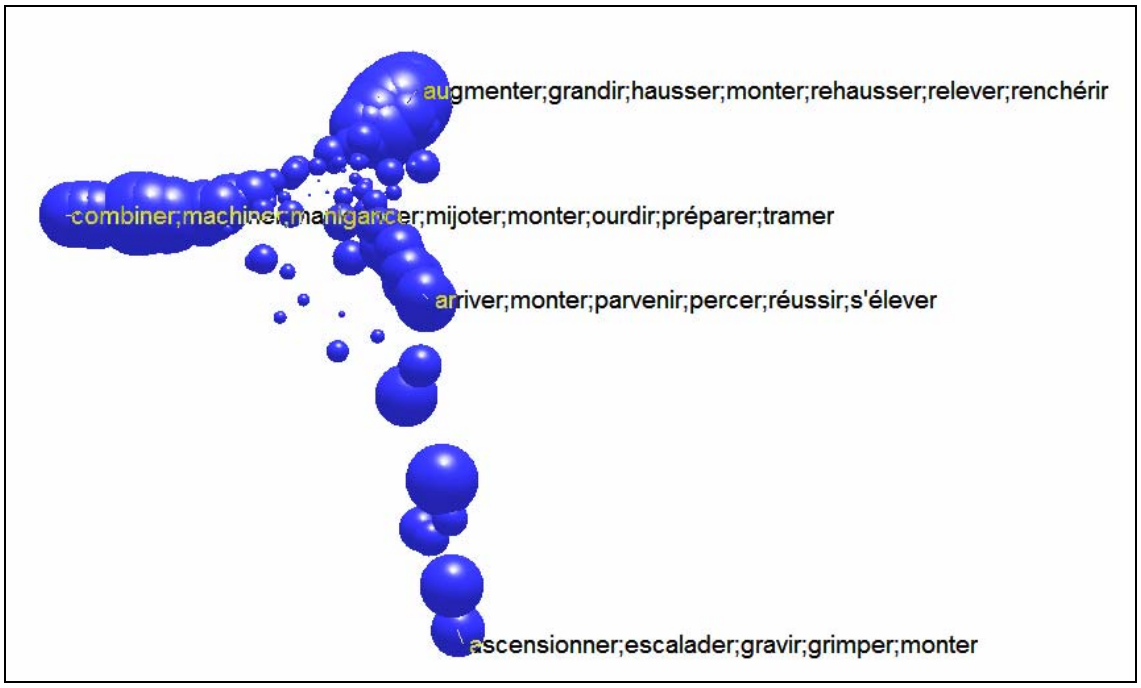
Rappelons que le modèle de la construction dynamique du sens, sur lequel nous basons notre approche, cherche à décrire et décomposer le processus de construction du sens, c'est-à-dire à être capable, à partir d'un énoncé composé d'unités linguistiques, de modéliser la construction du sens de l'énoncé ainsi que la construction du sens des unités qui le composent. Toute la difficulté étant de ne pas tomber dans un processus circulaire : le sens de l'énoncé est fonction du sens des unités qui le composent, et inversement le sens de ces unités dans cet énoncé est fonction du sens global de l'énoncé lui-même.

En désambiguïsant les noms par leur CSD et les verbes, au même titre que les adjectifs, par l'influence des CSD de noms sur leur sens, nous proposons un modèle permettant de casser ce processus circulaire de construction du sens.

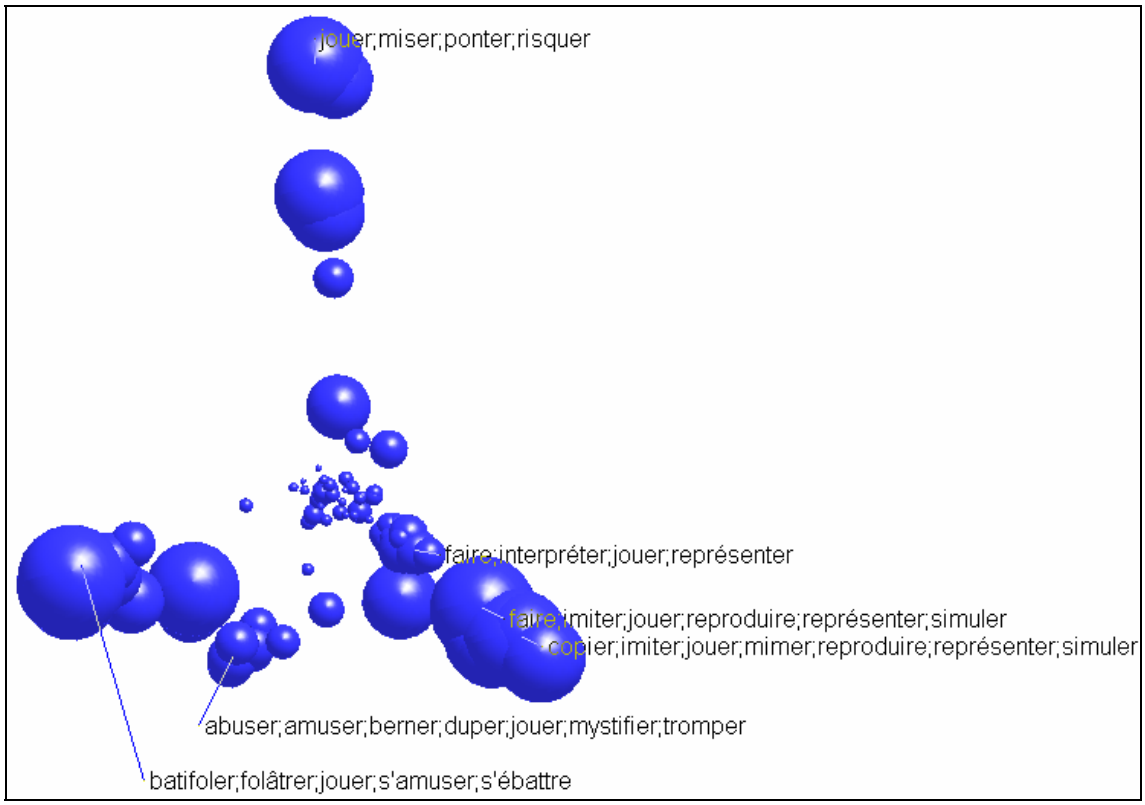
Annexes

A. Annexe 1 : visualisations 3D

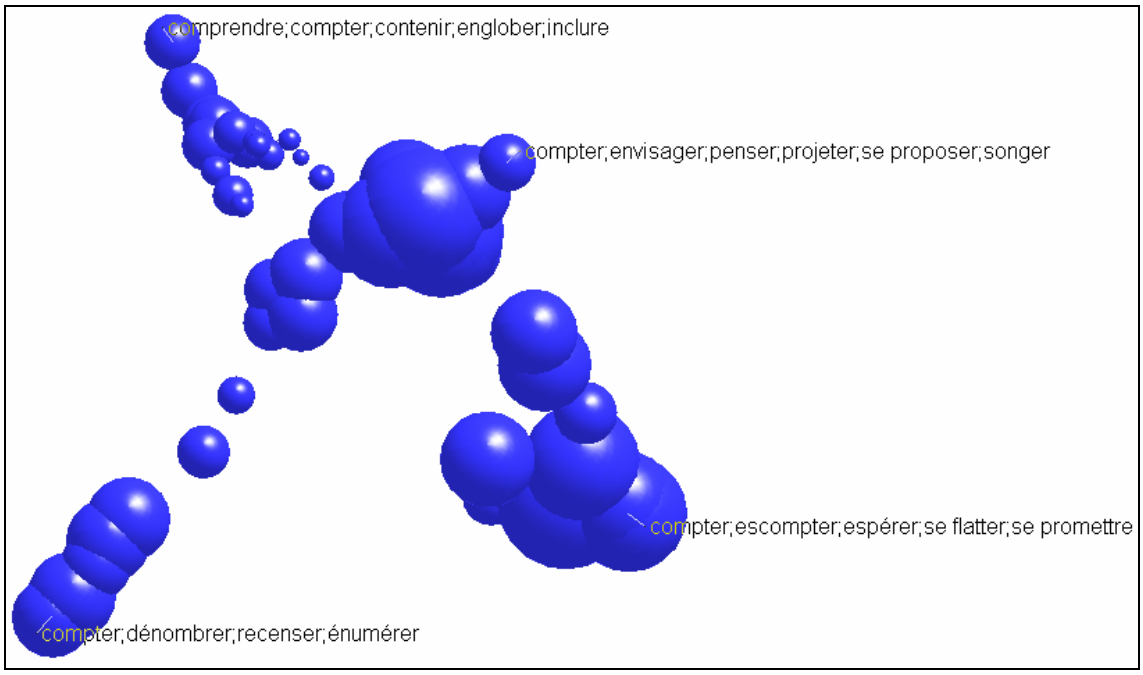
A.1. Visualisation 3D de l'espace sémantique du verbe monter



A.2. Visualisation 3D de l'espace sémantique du verbe **jouer**



A.3. Visualisation 3D de l'espace sémantique du verbe compter



B. Annexe 2 : extraits de code

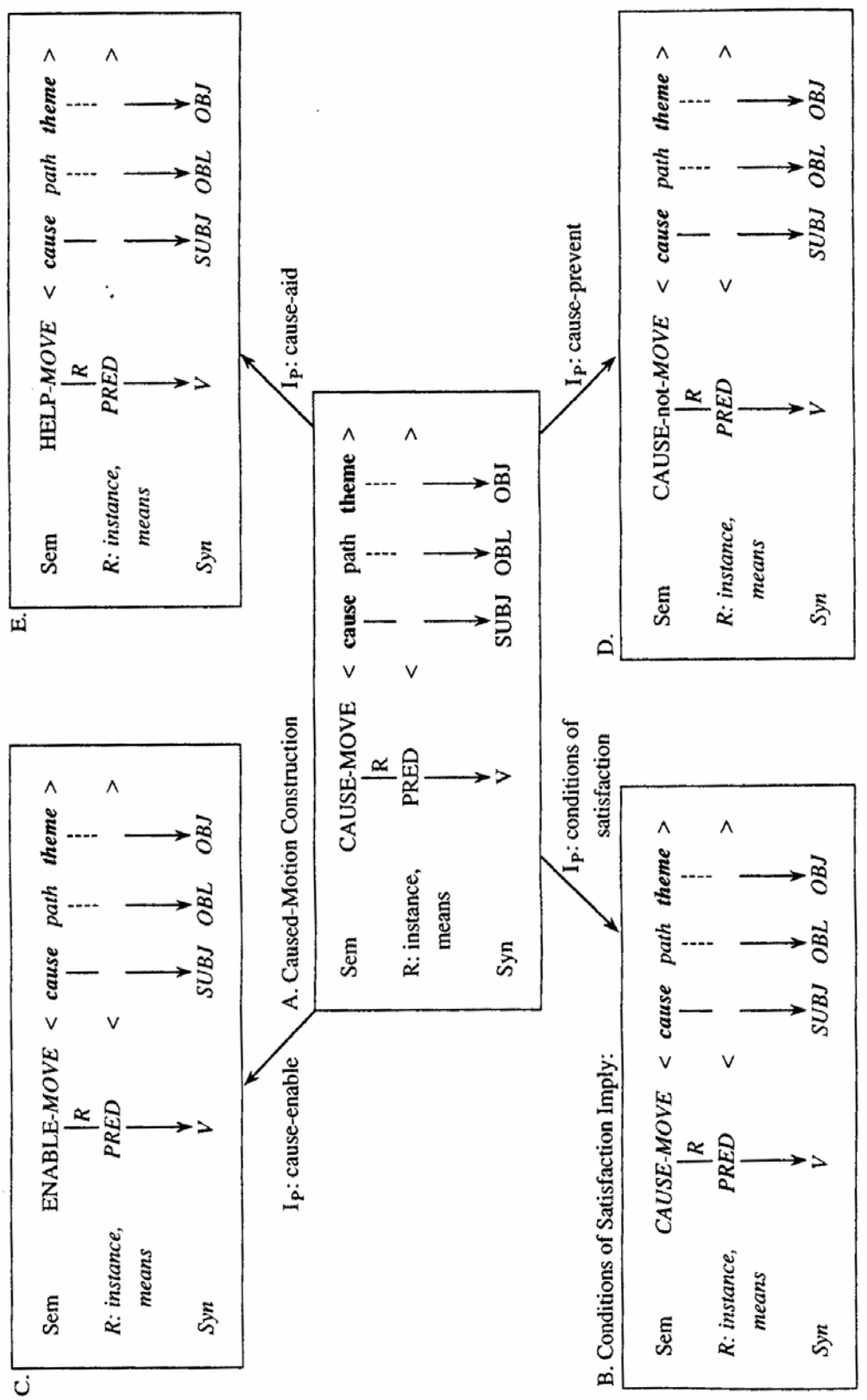
B.1. Mise en forme des cooccurrences, pour la construction de requête Frantext en Matlab

```
function motRequete = miseEnForme(mot,Type,ListeCodes)
    if (isempty(Type) == 1 | Type == [1])
        motRequete = mot;
    else
        if find(Type == 1)
            Type = Type(find(Type ~= 1));
        end
        if find(Type == 6)
            Type = union(Type(find(Type ~= 6)),[5 7 8 9]);
        end
        if (size(Type,2) == 1)
            catGram = strcat(' g=c',ListeCodes(Type(1)));
        else
            catGram = strcat(' g=(c',ListeCodes(Type(1)));
            for i=2:size(Type,2)
                catGram = strcat(catGram,'|c',ListeCodes(Type(i)));
            end
            catGram = strcat(catGram,')');
        end
        if find(Type == 4 | Type == 5 | Type == 6)
            if strncmp('s''',mot,2)
                mot = strrep(mot,'s''','');
                pronom = '(m''|t''|s'')';
            elseif strncmp('se ',mot,3)
                mot = strrep(mot,'se ','');
                pronom = '(me|te|se)';
            else
                pronom = '^(m''|t''|s''|me|te|se)';
            end
        else
            pronom = '';
        end
        if find(Type == 3 | Type == 11)
            graphie = strcat('c=@m',mot);
        else
            graphie = strcat('c=',mot);
        end
        motRequete = strcat(pronom,' @e(',graphie,catGram,')');
    end
```

B.2. requête Perl pour la construction compter sur son voisin *(les quatre mots sont mis en gras)*

```
$rep = "D:\\Guillaume\\Thèse\\Corpus\\";
@listefic = (
    $rep."select.anasynt");
$nbocc = 0;
$proportion =1;
foreach $nomfic (@listefic){
    open(CORPUS, $nomfic) || die "pas pu ouvrir ".$nomfic;
SEQUENCE : while(<CORPUS>){
if (/^<SEQ ([^<>]*)>/){
  $decision = rand;
  if ($decision<$proportion){
    $id = $1;
    $seq = $_;
    $texte = <CORPUS>;
    $analyse = <CORPUS>;
$i = 0;
$essai = 0;
pos $analyse = 0;
while ($analyse=~
/\t(V[^|]+|Pp[^|]+)\|compter\|[^|]+\|(\d+)\|([^|]+)\|/g){
  $pos1 = pos $analyse;
  $num1 = $2;
pos $analyse = 0;
while ($analyse =~
/\t(Prep)\|sur\|[^|]+\|(\d+)\|(([^|]+,)?(PREP|PRDE);$num1(,[^|]+)?)\|/g){
  $pos2 = pos $analyse;
  $num2 = $2;
pos $analyse = 0;
while ($analyse =~ /\t
(Nom[^|]+)\|voisin\|[^|]+\|(\d+)\|(([^|]+,)?NOMPREP;$num2(,[^|]+)?)\|/g){
  $pos4 = pos $analyse;
  $num4 = $2;
pos $analyse = 0;
while ($analyse =~
/\t(Det[^|]+)\|son\|[^|]+\|(\d+)\|(([^|]+,)?DET;$num4(,[^|]+)?)\|/g){
  $pos3 = pos $analyse;
  $num3 = $2;
$essai++;
if ($essai>1){
  $i++;
}
pos $analyse = $pos3;
}
pos $analyse = $pos4;
}
pos $analyse = $pos2;
}
pos $analyse = $pos1;
}
$i++;
if ($i == $essai){
    $nbocc+=$essai;
}
}
}
}
}
print $nbocc;
```

C. Annexe 3 : schéma des variantes de la construction de « mouvement induit » de Goldberg (1995 : 164)



D. Annexe 4 : exemple de sortie de Syntex

```
<SEQ idSeq=3;  idDoc=aussenac;>
<TXT>Toutes les difficultés liées à la construction de ces modèles sont cependant loin d' être résolues .
<ETIQ>Pro|tout|Toutes|1|0|0
      DetFP|le|les|2|DET;3|0
      NomFP|difficulté|difficultés|3|SUJ;11|DET;2,ADJ;4
      PpaFP|lier|liées|4|ADJ;3|PREP;5
      Prep|à|à|5|PREP;4|NOMPREP;7
      DetFS|le|la|6|DET;7|0
      NomFS|construction|construction|7|NOMPREP;5|DET;6,PRDE;8
      Prep|de|de|8|PRDE;7|NOMPREP;10
      DetMP|ce|ces|9|DET;10|0
      NomMP|modèle|modèles|10|NOMPREP;8|DET;9
      VCONJP|être|sont|11|0|SUJ;3
      Adv|cependant|cependant|12|0|0
      Prep|loin de|loin d'|13|0|NOMPREP;14
      VINFp|résoudre|être résolues|14|NOMPREP;13|0
      Typo|.|.|15|0|0
```

SEQ correspond à l'identifiant de l'énoncé

TXT contient l'énoncé analysé

ETIQ contient l'ensemble des éléments de l'énoncé analysé par Syntex. Pour chaque élément, voici le descriptif de l'analyse effectuée :

catégorie|lemme|mot|position|relations régi|relations recteur

E. Annexe 5 : extrait d'un questionnaire distribué pour l'évaluation psycholinguistique

Pour chaque énoncé, déterminez le sens du verbe ***jouer*** mis en gras.
Pour cela, attribuez une note à chaque synonyme en fonction de sa capacité à remplacer le sens du verbe *jouer*
Les différentes notes que l'on peut attribuer sont :

1 : pas synonyme
2 : peu synonyme
3 : assez synonyme
4 : très synonyme

Indication : il n'est pas nécessaire de forcer les notes pour avoir au moins un "4" parmis les huit synonymes

N°	Extrait complet	miser au sens de parier, risquer	s'amuser au sens de rire, se divertir	imiter au sens de mimer, reproduire, simuler	influer au sens de agir	se mouvoir au sens de remuer, rouler	interpréter au sens de représenter, incarner	abuser au sens de duper, rouler, tromper	pratiquer au sens de manier, tenir
Exemple	nous nous retrouvons souvent pour procéder aux échanges, on y trouve des retraités niçois et surtout des soldats italiens, les copains de Marcello qui chantent l'opéra et **<*jouent*>** de la guitare avant de monter des gardes folkloriques dans les endroits stratégiques de la ville. Nous y voilà, c'est tout petit, la mère Rosso laisse toujours la porte de		1						3
3	, avec les cris de ses baigneurs, avec ses tentes orange et rouge qui claquaient dru, une petite fanfare attardée, mais encore alerte et pleine de feu, qui continue à **<*jouer*>** après la clôture de la fête, et d'un coup il se sentit rassuré ; la saison reflambait encore, un peu miraculeusement, au bout de cette longue roule étouffée sous les housses								
4	... - Tant pis, jeta Jami, et il fit dérailler son train. Après un temps, Olivier dit sur un ton affecté : - Oh! et puis, après tout, si tu veux. Mais assois-toi sur le canapé. Je **<*jouerai*>** et tu regarderas. Avec des contorsions de visage, Olivier parvint à se faire la tête de Michel Simon, celle de Fernandel, prit l'attitude du vieux noble d'André Lefaur.								
7	il pensait ; et *Christophe en éprouvait un bien-être inexprimable. *Hassler ne songeait plus à calculer le nombre des pages qui étaient jouées et celles qui restaient à **<*jouer*>**. Quand *Christophe avait fini un morceau, il disait : -après ! ... après ! ... il commençait à faire usage du langage humain. -bon, cela ! Bon ! ... (s' exclamait-il). Fameux								
8	de *Diriclet. Elles ont permis de donner une forme intuitive aux théorèmes d' existence des solutions des équations différentielles ou aux dérivées partielles. Elles **<*jouent*>** un rôle utile dans l' étude des fonctions presque périodiques de *M *Harald *Bohr. Soit *F / *X /, une fonction presque périodique de la variable réelle *X, définie et continue								
10	. - On s'assied ? fait Armoire. Et nous voilà tous trois assis en cercle autour d'une caisse fermée sur laquelle Asperge a posé la bouteille. - J'y pense, dit Asperge, si on **<*jouait*>** à "l'avant-dernier- qui-boit" ? - à quoi ? demande Armoire. C'est bien plus tard que j'ai compris combien leur petit numéro était au point. Salauds. - Mais oui, dit A' tu								

F. Annexe 6 : WordNet

WordNet (Fellbaum 1998) est une base de données lexicale développée à l'université Princeton. Son but est de répertorier, classifier et relier de diverses manières le contenu sémantique et lexical de la langue anglaise. Le système se présente sous la forme d'une base de données électronique accessible gratuitement. La composante atomique sur laquelle repose le système entier est le synset (synonym set), c'est-à-dire un groupe de mots interchangeables, dénotant un sens ou un usage particulier. La version 2.1 de WordNet contient 155 327 mots, 117 597 synsets et 207 016 sens. Par exemple, le nom *car* contient cinq synsets dénotant chacun un sens différent, décrit par une courte définition :

- S: (n) car, auto, automobile, machine, motorcar (a motor vehicle with four wheels; usually propelled by an internal combustion engine) *"he needs a car to get to work"*
- S: (n) car, railcar, railway car, railroad car (a wheeled vehicle adapted to the rails of railroad) *"three cars had jumped the rails"*
- S: (n) cable car, car (a conveyance for passengers or freight on a cable railway) *"they took a cable car to the top of the mountain"*
- S: (n) car, gondola (the compartment that is suspended from an airship and that carries personnel and the cargo and the power plant)
- S: (n) car, elevator car (where passengers ride up and down) *"the car was on the top floor"*

Les synsets peuvent aussi représenter des concepts plus abstraits, de plus haut niveau que les mots et leurs sens, qu'on peut organiser sous forme d'ontologies. WordNet définit ainsi des relations sémantiques permettant d'organiser le sens des mots (et donc par extension les mots eux-mêmes). Les types de relations sémantiques sont les suivants : hyperonymes, hyponymes, troponymes (is a troponym of **Y** if **to X** is **to Y** in some manner), forme dérivée liée, structure actantielle (sentence frame), domaine. On pourra par exemple interroger le système quant aux hyperonymes d'un mot particulier : à partir du sens le plus commun du nom *car* (le sens *car, auto*…), la relation d'hyperonymie définit un arbre de concepts de plus en plus généraux (cf. figure page suivante). Notons le développement considérable de la hiérarchie proposée par WordNet, en comparant les relations hyperonymiques actuelles (1ère structure) avec celles proposées il y a simplement deux ans (2ème structure).

- S: (n) car, auto, automobile, machine, motorcar (a motor vehicle with four wheels; usually propelled by an internal combustion engine) *"he needs a car to get to work"*
  - S: (n) motor vehicle, automotive vehicle (a self-propelled wheeled vehicle that does not run on rails)
    - S: (n) self-propelled vehicle (a wheeled vehicle that carries in itself a means of propulsion)
      - S: (n) wheeled vehicle (a vehicle that moves on wheels and usually has a container for transporting things or people) *"the oldest known wheeled vehicles were found in Sumer and Syria and date from around 3500 BC"*
        - S: (n) vehicle (a conveyance that transports people or objects)
          - S: (n) conveyance, transport (something that serves as a means of transportation)
            - S: (n) instrumentality, instrumentation (an artifact (or system of artifacts) that is instrumental in accomplishing some end)
              - S: (n) artifact, artefact (a man-made object taken as a whole)
                - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) *"how big is that part compared to the whole?"; "the team is a unit"*
                  - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
                    - S: (n) physical entity (an entity that has physical existence)
                      - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
        - S: (n) container (any object that can be used to hold things (especially a large metal boxlike object of standardized dimensions that can be loaded from one form of transport to another))
          - S: (n) instrumentality, instrumentation (an artifact (or system of artifacts) that is instrumental in accomplishing some end)
            - S: (n) artifact, artefact (a man-made object taken as a whole)
              - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) *"how big is that part compared to the whole?"; "the team is a unit"*
                - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
                  - S: (n) physical entity (an entity that has physical existence)
                    - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

```
1. car, auto, automobile, machine, motorcar
    => motor vehicle, automotive vehicle
        => vehicle
            => conveyance, transport
                => instrumentality, instrumentation
                    => artifact, artefact
                        => object, physical object
                            => entity, something
```

G. Annexe 7 : listes de cliques

G.1. Cliques contenant le verbe monter

1 : affluer, aller, arriver, monter
2 : affluer, augmenter, monter
3 : agencer, ajuster, bâtir, monter
4 : agencer, ajuster, combiner, disposer, monter, organiser
5 : agencer, appareiller, disposer, monter
6 : agencer, bâtir, fabriquer, monter
7 : agencer, combiner, fabriquer, monter, tisser
8 : agencer, combiner, manigancer, monter, ourdir, tisser
9 : agencer, disposer, installer, monter, organiser
10 : ajuster, assembler, bâtir, monter
11 : ajuster, assembler, combiner, disposer, monter
12 : ajuster, monter, remonter
13 : aller, arriver, atteindre, monter, s'élever
14 : aller, arriver, entrer, monter
15 : aller, atteindre, monter, prendre
16 : aller, entrer, monter, prendre
17 : aller, monter, prendre, voler
18 : aller, monter, progresser
19 : aller, monter, s'élever, voler
20 : appareiller, assembler, disposer, monter
21 : appareiller, disposer, monter, préparer
22 : appareiller, monter, s'accoupler
23 : arriver, atteindre, monter, parvenir, percer, s'élever
24 : arriver, entrer, monter, parvenir
25 : arriver, monter, parvenir, percer, réussir, s'élever
26 : ascensionner, escalader, gravir, grimper, monter
27 : assembler, bâtir, coudre, monter
28 : assembler, combiner, monter, tresser
29 : assembler, enchâsser, monter, sertir
30 : assembler, lever, monter
31 : atteindre, gravir, monter
32 : atteindre, monter, porter
33 : atteindre, monter, réaliser
34 : atteindre, monter, s'élever, se monter
35 : augmenter, aviver, exciter, monter
36 : augmenter, aviver, monter, rehausser
37 : augmenter, croître, grandir, grossir, monter, s'accroître
38 : augmenter, croître, grandir, grossir, monter, s'intensifier
39 : augmenter, croître, grandir, monter, s'accroître, s'élever
40 : augmenter, doubler, monter, redoubler
41 : augmenter, exciter, monter, relever
42 : augmenter, exhausser, hausser, lever, monter, relever, élever
43 : augmenter, exhausser, hausser, monter, relever, remonter, élever
44 : augmenter, exhausser, hausser, monter, surhausser, élever
45 : augmenter, forcer, grossir, hausser, monter, renchérir
46 : augmenter, grandir, grossir, hausser, monter, renchérir
47 : augmenter, grandir, grossir, monter, s'amplifier, s'intensifier
48 : augmenter, grandir, hausser, monter, rehausser, relever, renchérir
49 : augmenter, grandir, hausser, monter, rehausser, relever, élever
50 : augmenter, grandir, monter, redoubler
51 : augmenter, grossir, hausser, majorer, monter, renchérir
52 : augmenter, hausser, lever, monter, rehausser, relever, élever

53 : augmenter, hausser, majorer, monter, rehausser, relever, renchérir, revaloriser
54 : augmenter, hausser, majorer, monter, rehausser, relever, revaloriser, élever
55 : augmenter, hausser, monter, rehausser, relever, remonter, renchérir, revaloriser
56 : augmenter, hausser, monter, rehausser, relever, remonter, revaloriser, élever
57 : augmenter, hausser, monter, rehausser, surhausser, élever
58 : augmenter, monter, s'accentuer
59 : bâtir, constituer, créer, faire, monter, établir
60 : bâtir, créer, fabriquer, faire, monter
61 : bâtir, créer, faire, monter, élever, établir
62 : bâtir, monter, échafauder, établir
63 : chevaucher, couvrir, monter
64 : combiner, disposer, monter, organiser, préparer
65 : combiner, fabriquer, monter, préparer, tisser
66 : combiner, machiner, manigancer, mijoter, monter, ourdir, préparer, tramer
67 : combiner, machiner, manigancer, monter, nouer, ourdir, préparer, tisser, tramer, tresser
68 : combiner, monter, organiser, préparer, tramer
69 : constituer, créer, monter, organiser, établir
70 : constituer, disposer, monter, organiser, établir
71 : couvrir, disposer, monter
72 : couvrir, faire, monter, servir
73 : couvrir, franchir, monter
74 : couvrir, monter, s'accoupler, saillir, servir
75 : croître, monter, progresser, s'accroître
76 : créer, fabriquer, faire, monter, réaliser
77 : créer, faire, jouer, monter
78 : créer, faire, monter, soulever, élever
79 : créer, monter, porter, élever
80 : disposer, dresser, installer, monter, établir
81 : disposer, dresser, monter, préparer
82 : disposer, installer, monter, organiser, établir
83 : disposer, monter, prendre
84 : doubler, franchir, monter
85 : dresser, exciter, monter, relever, soulever
86 : dresser, faire, lever, monter, soulever, élever
87 : dresser, faire, monter, préparer
88 : dresser, faire, monter, élever, établir
89 : dresser, hausser, lever, monter, relever, soulever, élever
90 : dresser, monter, planter, élever
91 : embarquer, enlever, monter, prendre
92 : embarquer, monter, porter
93 : enchatonner, enchâsser, monter, sertir
94 : enfourcher, monter, percer
95 : enlever, exciter, monter, soulever
96 : enlever, forcer, monter, prendre
97 : enlever, hisser, lever, monter, soulever, élever
98 : enlever, jouer, monter
99 : enlever, monter, prendre, soulever
100 : enlever, monter, prendre, voler
101 : entrer, forcer, monter, prendre
102 : entrer, monter, planter
103 : entrer, monter, s'engouffrer
104 : entrer, monter, s'installer
105 : escalader, franchir, gravir, monter
106 : escalader, grimper, monter, se hisser
107 : exciter, monter, porter, relever
108 : exhausser, hausser, monter, relever, surélever, élever
109 : exhausser, hausser, monter, surhausser, surélever, élever
110 : fabriquer, faire, monter, préparer

111 : faire, mijoter, monter, préparer
112 : faire, monter, prendre, réussir
113 : faire, monter, prendre, soulever
114 : faire, monter, procurer, soulever
115 : grimper, monter, s'élever, se hisser
116 : grimper, monter, se percher
117 : hausser, hisser, lever, monter, soulever, élever
118 : hausser, hisser, monter, porter, élever
119 : hausser, monter, porter, relever, élever
120 : hausser, monter, rehausser, relever, surélever, élever
121 : hausser, monter, rehausser, surhausser, surélever, élever
122 : hausser, monter, relever, remonter, soulever, élever
123 : hausser, monter, soulever, surhausser, élever
124 : jouer, mettre en scène, monter
125 : mettre en scène, monter, préparer
126 : mijoter, mitonner, monter, préparer
127 : mijoter, monter, préparer, échafauder
128 : monter, nouer, établir
129 : monter, parvenir, se hausser
130 : monter, pourvoir, procurer
131 : monter, pourvoir, établir
132 : monter, s'embarquer, s'engouffrer
133 : monter, s'envoler, s'échapper
134 : monter, s'envoler, s'élever, voler
135 : monter, s'édifier, se construire
136 : monter, s'élever, se bâtir, se monter
137 : monter, se bâtir, se construire
138 : monter, se débourgeoiser
139 : monter, se débourrer
140 : monter, se guinder, se hausser, se hisser

G.2. Cliques contenant le verbe **jouer**

1 : abuser, agir, jouer
2 : abuser, amuser, berner, duper, jouer, mystifier, tromper
3 : abuser, berner, duper, jouer, mystifier, refaire, rouler, tromper
4 : abuser, feindre, jouer, mentir, tromper
5 : abuser, jouer, mentir, mystifier, tromper
6 : abuser, jouer, s'amuser
7 : affecter, attaquer, jouer, toucher
8 : affecter, contrefaire, feindre, jouer, simuler
9 : affecter, contrefaire, jouer, simuler, singer
10 : affecter, jouer, plastronner, poser
11 : affecter, jouer, poser, simuler
12 : agir, exécuter, faire, jouer
13 : agir, faire, jouer, représenter
14 : agir, faire, jouer, user
15 : agir, fonctionner, jouer
16 : agir, influer, jouer
17 : agir, intervenir, jouer
18 : agiter, jouer, manier
19 : agiter, jouer, remuer, secouer
20 : amuser, berner, duper, flouer, jouer, mystifier, tromper
21 : amuser, jouer, rire
22 : aventurer, compromettre, exposer, hasarder, jouer, risquer
23 : avoir du jeu, jouer
24 : badiner, blaguer, jouer, plaisanter, railler, rire
25 : badiner, blaguer, jouer, plaisanter, taquiner
26 : badiner, folâtrer, jouer, plaisanter, s'amuser
27 : badiner, jouer, plaisanter, railler, rire, s'amuser
28 : badiner, jouer, plaisanter, s'amuser, taquiner
29 : batifoler, folâtrer, jouer, plaisanter, s'amuser
30 : batifoler, folâtrer, jouer, s'amuser, s'ébattre
31 : berner, duper, flouer, jouer, mystifier, refaire, rouler, tromper
32 : berner, jouer, railler
33 : blaguer, jouer, mentir, plaisanter
34 : boursicoter, hasarder, jouer
35 : boursicoter, jouer, miser
36 : boursicoter, jouer, spéculer, tripoter
37 : contrefaire, copier, imiter, jouer, mimer, reproduire, simuler, singer
38 : contrefaire, faire, imiter, jouer, reproduire, simuler, singer
39 : contrefaire, feindre, imiter, jouer, simuler
40 : copier, imiter, jouer, mimer, reproduire, représenter, simuler
41 : coulisser, jouer
42 : créer, faire, interpréter, jouer
43 : créer, faire, jouer, monter
44 : donner, exposer, jouer, représenter
45 : donner, faire, jouer, représenter
46 : donner, intervenir, jouer
47 : donner, jouer, passer
48 : duper, faire, jouer, refaire
49 : enlever, gratter, jouer, racler
50 : enlever, jouer, monter
51 : enlever, jouer, souffler
52 : exposer, jouer, tourner
53 : exécuter, faire, interpréter, jouer
54 : exécuter, faire, jouer, pratiquer, tenir
55 : faire, figurer, imiter, jouer, reproduire, représenter

56 : faire, imiter, jouer, reproduire, représenter, simuler
57 : faire, interpréter, jouer, représenter
58 : faire, jouer, pratiquer, user
59 : faire, jouer, refaire, reproduire
60 : feinter, jouer, mystifier, rouler, tromper
61 : figurer, incarner, jouer, reproduire, représenter
62 : flamber, jouer, refaire
63 : folâtrer, jouer, s'ébattre, s'ébrouer
64 : fonctionner, jouer, marcher, se mouvoir
65 : fonctionner, jouer, marcher, tourner
66 : fonctionner, jouer, remuer, se mouvoir
67 : fonctionner, jouer, remuer, tourner
68 : gauchir, gondoler, jouer
69 : gratter, jouer, passer
70 : gratter, jouer, remuer
71 : incarner, interpréter, jouer, représenter
72 : influer, jouer, tourner
73 : interpréter, jouer, tourner
74 : jongler, jouer
75 : jouer, manier, pratiquer, tenir
76 : jouer, manier, pratiquer, user
77 : jouer, manier, toucher, tripoter
78 : jouer, marcher, passer, tourner
79 : jouer, marcher, rouler, se mouvoir
80 : jouer, marcher, rouler, tourner
81 : jouer, mettre en scène, monter
82 : jouer, mettre en scène, représenter
83 : jouer, miser, parier, ponter
84 : jouer, miser, ponter, risquer
85 : jouer, pianoter
86 : jouer, pincer
87 : jouer, pratiquer, s'entraîner
88 : jouer, remuer, rouler, se mouvoir
89 : jouer, remuer, rouler, toucher
90 : jouer, remuer, rouler, tourner
91 : jouer, rire, s'amuser, se divertir
92 : jouer, s'amuser, s'ébattre, se divertir
93 : jouer, s'entraîner, s'exercer
94 : jouer, s'ébrouer, souffler
95 : jouer, se faire entendre, sonner
96 : jouer, sonner, souffler

G.3. Cliques contenant le verbe **compter**

1 : apprécier, calculer, compter, estimer, mesurer, peser, évaluer
2 : apprécier, comprendre, compter, entendre
3 : apprécier, compter, considérer, croire, estimer
4 : apprécier, compter, considérer, estimer, examiner, peser
5 : apprécier, compter, estimer, examiner, peser, évaluer
6 : attendre, compter, croire, espérer, présumer
7 : attendre, compter, escompter, espérer, se promettre
8 : avoir l'intention, compter, penser
9 : avoir pour certain, compter
10 : avoir pour sûr, compter
11 : calculer, chiffrer, compter, supputer, évaluer
12 : calculer, compter, estimer, mesurer, peser, supputer, évaluer
13 : comprendre, compter, contenir, englober, inclure
14 : comprendre, compter, prendre
15 : compter, computer
16 : compter, considérer, croire, estimer, penser
17 : compter, considérer, croire, estimer, réputer
18 : compter, considérer, envisager, estimer, examiner, penser, regarder
19 : compter, considérer, envisager, estimer, prendre, regarder
20 : compter, considérer, envisager, penser, songer
21 : compter, considérer, estimer, examiner, penser, peser, regarder
22 : compter, considérer, estimer, prendre, regarder, réputer
23 : compter, considérer, estimer, regarder, réputer, tenir pour
24 : compter, contenir, mesurer
25 : compter, contenir, posséder
26 : compter, contenir, présenter
27 : compter, croire, espérer, penser, présumer, s'attendre
28 : compter, croire, estimer, penser, présumer
29 : compter, croire, estimer, présumer, réputer
30 : compter, croire, penser, présumer, supposer
31 : compter, dater, marquer
32 : compter, décompter, dénombrer
33 : compter, dénombrer, inventorier, énumérer
34 : compter, dénombrer, nombrer, énumérer
35 : compter, dénombrer, nombrer, évaluer
36 : compter, dénombrer, recenser, énumérer
37 : compter, dénombrer, recenser, évaluer
38 : compter, entendre, posséder
39 : compter, entrer en ligne de compte, importer, peser
40 : compter, envisager, penser, projeter, se proposer, songer
41 : compter, escompter, espérer, s'attendre
42 : compter, escompter, espérer, se flatter, se promettre
43 : compter, escompter, espérer, supputer
44 : compter, escompter, espérer, tabler
45 : compter, espérer, penser, se flatter, se targuer
46 : compter, estimer, examiner, peser, supputer, évaluer
47 : compter, estimer, penser, présumer, regarder
48 : compter, estimer, présumer, regarder, réputer, tenir pour
49 : compter, examiner, inventorier, regarder
50 : compter, exister
51 : compter, facturer
52 : compter, fonder, tabler
53 : compter, importer, introduire
54 : compter, importer, être important
55 : compter, inclure, introduire

56 : compter, introduire, présenter
57 : compter, marquer, présenter
58 : compter, marquer, supposer
59 : compter, nombrer, supputer, évaluer
60 : compter, payer
61 : compter, posséder, prendre
62 : compter, prendre, supposer
63 : compter, précompter
64 : compter, s'appuyer, tabler
65 : compter, se ranger
66 : compter, épargner

Index

actant, 37, 143, 150

AFC, 97, 135, 204, 210, 213

ambiguïté, 13, 31, 48, 67-70, 75, 94, 105-112, 135, 155, 171, 180, 183, 199, 206, 222-226

analyseur syntaxique, 82, 120, 155, 160, 178, 184, 187, 198, 216, 219, 223

annotateur, 49

apprentissage supervisé, 44-47, 54-56, 68, 74, 79

approche

- **endogène**, 44, 67, 84, 186, 226
- **exogène**, 43, 49
- **mixte**, 44, 54, 66, 84

arbre de décision, 59

bassin, 107, 111, 179

borne, 56, 66, 73-76

bruit, 12, 40, 113, 128, 184, 198, 218, 221

classe sémantique, 38, 45, 54, 62, 67, 79, 83-87, 142, 186, 199-212, 218, 224-226

clique, 96, 99-104, 112-117, 125-136, 156, 159, 165, 173, 205, 213, 239

clusterisation, 69, 206, 210, 220

cognition, 16, 32, 141

collocation, 49, 57, 71, 79, 81, 218

complément essentiel/circonstanciel, 91, 125, 128, 151, 160, 179, 184-199, 202, 205, 216, 223

compositionnalité, 106, 146, 218

continu, 18, 23, 24, 86, 90-94, 101, 135, 138, 186

CSD, 199-211, 218, 223-227, 228

CTRW, 57, 60

degré d'affinité, 112-117, 122, 125, 127-135, 155, 159, 180, 213

degré de fiabilité, 56, 67, 180, 213

ditransitive, 142, 144, 146, 149, 217

espace distributionnel, 201-205, 219

espace sémantique, 17, 20, 84, 86, 95-104, 107, 109, 113, 117, 121, 125, 135, 157-163, 169, 173, 179, 199, 205, 213, 215, 229-231

essentialiste, 31

étiqueteur, 13, 54, 74, 223

étude psycholinguistique, 16, 136, 172, 174-176, 236

facette, 29, 130, 210

Frantext, 119, 122-133, 160, 169, 210, 214, 232

gestaltiste, 93, 106

Goldberg, 39, 137-153, 217, 234

granularité, 71-76, 111

graphe, 95, 116

homonymie, 22-25, 53, 70-73

hyperonyme, 27, 59, 63, 72, 226-228, 237

hyponyme, 27, 59, 228, 237

indétermination, 61, 94, 104, 107-111, 155, 169, 189

indexation, 12, 42, 77, 195, 219

information mutuelle, 51, 115, 125, 132, 191-193, 213

interjuge, 61, 74, 111

Kay, 137, 152

Kleiber, 22, 28-32

LDOCE, 54-57, 73, 82

Le monde

LM3, 120, 155, 160, 175, 214

LM10, 120, 202, 214

lexique grammaire, 137, 139

log-vraisemblance, 48

LSA, 45, 68, 75, 95

métaphore, 25-28

métonymie, 24-29

monosémie, 23, 189

nominaliste, 31

noyau de sens, 31-34, 150

paraphrase, 17-20, 34, 212, 219, 223

pattern, 11, 46-48, 71, 226

Perl, 161, 233

polytaxie, 38

précision, 48, 61, 73, 85, 136, 162, 172, 218

préposition, 144, 151, 161, 168, 187

productivité, 15, 186

rattachement prépositionnel, 184, 189, 223

région sémantique, 94, 97, 103-113, 125-135, 164-167, 180, 185, 199, 210, 214, 217

r**elation co-textuelle**, 130, 133, 135

relation de dépendance, 120, 156, 184

réseau, 44, 50-53, 79

r**ôle actanciel**, 143-146, 150

Schütze, 45, 60, 67-70, 72, 77

sémantaxe, 38, 217

sens

d'un énoncé, 15-17, 34, 153, 212, 228

d'un mot, 16, 23, 42, 50, 55, 65, 69, 71-73, 80, 94, 106

d'une construction, 15, 20

Senseval, 71, 74-79

source de connaissances, 20, 45, 54, 57, 68, 79, 84

SQL, 163

synset, 56, 59-65, 76, 237

Syntex, 14, 120, 155, 156, 162, 184, 187, 198-202, 219-223, 235

système dynamique, 94

t**axonomie**, 62, 80

thésaurus, 43, 62, 85

TLFi, 17, 33, 85, 87-91, 99, 110, 126

traduction automatique, 42, 58, 222

transitif, 80, 88-91, 110, 160, 172, 178

triplet, 120, 187, 202, 214, 219

t-score, 213

Visusyn, 122, 176, 199, 211

way-construction, 40

Web, 50, 79, 119, 122, 188, 196, 214, 222

WordNet, 43, 53, 56-60, 62, 66, 75-77, 86, 189, 237

Zipf, 214, 280

zone de sens, 86, 134, 167, 170, 174

Bibliographie

AGIRRE E., MARTINEZ D. (2000). Exploring automatic word sense disambiguation with decision lists and the web. *Workshop on Semantic Annotation and Intelligent Content (COLING-2000).*

AGIRRE E., MARTINEZ D. (2001). Knowledge sources for word sense disambiguation. *4th International Conference on Text Speech and Dialogue*, pp. 1-10.

AGIRRE E., MARTINEZ D. (2004). Unsupervised WSD based on automatically retrieved examples: The importance of bias. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'2004)*, Barcelona, Spain.

AUDIBERT L. (2002). Etude des critères de désambiguïsation sémantique automatique : Présentation et premiers résultats sur les cooccurrences. *6ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL-2002)*, pp. 415-424.

AUDIBERT L. (2003). *Outils d'exploration de corpus et désambiguïsation lexicale automatique*. Thèse de doctorat (Informatique), Équipe DEscription Linguistique Informatisée sur Corpus (DELIC), Université d'Aix-Marseille I - Université de Provence, Aix-en-Provence.

AUDIBERT L. (2003). *Etude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences*, 10ème conférence sur le Traitement Automatique des Langues Naturelles (TALN-2003), Batz-sur-Mer, pp. 35-44.

AUSSENAC-GILLES N., BIÉBOW B., SZULMAN N. (2000). Revisiting Ontology Design: a method based on corpus analysis, Actes de 12th International Conference on Knowledge Engineering and Knowledge Management. Juan-Les-Pins

BOOKMAN L. (1987). A microfeature based scheme for modelling semantics, *10th International Joint Conference on Artificial Intelligence (IJCAI-1987)*, pp. 611-614.

BOURIGAULT D. (1993). « Analyse syntaxique locale pour le repérage de termes complexes dans un texte », *TAL*, n°34-2.

BOURIGAULT D., FABRE C. (2000). Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaires,* n° 25, Université Toulouse - Le Mirail, pp. 131-151.

BOURIGAULT D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, Actes de *TALN 2002*, Nancy, pp. 75-84

BOURIGAULT D., FREROT C. (2004). *Ambiguïté de rattachement prépositionnel : introduction de ressources exogènes de sous-catégorisation dans un analyseur syntaxique de corpus endogène.* Actes de la Conférence TALN, Fès, 19-22 avril.

BOURIGAULT D., FREROT C. (2005). Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. Actes de la Conférence TALN, Dourdan, 6-10 juin.

BRUCE R., GUTHRIE. L. (1992). Genus disambiguation: A study in weighted preference. In *Proceedings of COLING-92*, Nantes, pp. 1187-1191.

CADIOT P., BERTHONNEAU A.-M. (1993). *Les prépositions, méthode d'analyse*, presse Universitaire de Lille.

CADIOT P. (1997). *Les prépositions abstraites en français*, Paris, Armand Colin.

CADIOT P. (1999). Les sens de *jouer* : esquisse d'une approche par le biais des attaches prépositionnelles, *Recherches en linguistique et psychologie cognitive*, n°11, presses Universitaires de Reims.

CADIOT P., VISETTI P.Y. (2001). *Pour une théorie des formes sémantiques, Motifs, profils, thèmes*, Paris, PUF.

CHAFE, W. L. (1970). *Meaning and the Structure of Language*. Chicago, The University of Chicago Press.

CHURCH K. W., HANKS P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Machine Translation*, vol. 16, n°1, pp. 22-29.

COOK, W. A. (1989). *Case Grammar Theory*. Washington, DC, Georgetown University Press.

CROFT W., CRUSE D.A. (2004). *Cognitive Linguistics*, Cambridge University Press.

CRUSE D.A. (2000). « Aspects of the microstructure of word meanings », in RAVIN Y., LEACOCK C. (eds), *Polysemy: theoretical and computationl approaches*, Oxford University Press, pp. 30-51.

CULIOLI A. (1990). *Pour une linguistique de l'énonciation*, Paris, Ophrys.

DEERWESTER S.C., DUMAIS S. T., LANDAUER T. K., FURNAS G. W., HARSHMAN R. A. (1990). Indexing by Latent Semantic Analysis. JASIS 41(6), pp. 391-407.

DESCLES J.P. (1985). Représentation des connaissances : archétypes cognitifs, schèmes conceptuels, schèmes grammaticaux, *Actes Sémiotiques*, Documents(VII), pp. 69-70.

DESCLES J.P. (1990). *Langages Applicatifs, Langues Naturelles et Cognition*. Hermès, Paris.

DIAB M. (2003). Word sense disambiguation within a multilingual framework. *PhD Thesis*, University of Maryland, College Park.

DINI L., TOMASO V. DI, SEGOND F. (1998). Error Driven Word Sense Disambiguation, COLING-ACL 1998, pp. 320-324.

DUVIGNAU K. (2002). La métaphore, berceau et enfant de la langue. La métaphore verbale comme approximation sémantique par analogie dans les textes scientifiques et les productions enfantines (2-4 ans), Thèse en Sciences du Langage, Université Toulouse Le-Mirail.

DUVIGNAU, K. (2003). « Métaphore verbale et approximation », in DUVIGNAU, K., GASQUET, O., GAUME, O. (eds) *Regards croisés sur l'analogie*. *Revue d'Intelligence Artificielle*, spécial, Vol 5/6. Hermès Sciences, pp. 869-881.

ENJALBERT P. (2005), *Semantique et TALN*, Paris, Hermès.

ESCUDERO G., MARQUEZ L., RIGAU, G. (2000). Boosting applied to word sense disambiguation. *11th European Conference on Machine Learning (ECML-2000)*, pp. 129-141.

FABRE C., FRÉROT C. (2002). Groupes prépositionnels arguments ou circonstants : vers un repérage automatique en corpus, *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002)*, Nancy.

FAUCONNIER G. (1984). *Les espaces mentaux*, Editions de Minuit.

FELLBAUM C. (1998). *Wordnet: an Electronic Lexical Database*, Cambridge, MIT Press.

FILLMORE C. (1971). Some Problems of Case Grammar, In O'BRIAN R., ed., *Report of the twenty-Second annual Round Table Meeting on Languages and Linguistics.* Washington, Georgetown University Press.

FILLMORE C., KAY P., O'CONNOR C. (1988). Regularity and Idiomaticity in Grammatical Constructions: The case of *Let Alone*, Language n°64, pp. 501-538.

FLEURY S. (1998), Gaspar, un dispositif de TALN basé sur la programmation à Prototypes, Actes de TALN'98, Paris.

FRANÇOIS J., MANGUIN J.L. VICTORRI B. (2003). La réduction de la polysémie adjectivale en co-texte nominal: une méthode de sémantique calculatoire, *Cahiers du Crisco, n°*14, http://www.crisco.unicaen.fr.

FRANÇOIS J. (2003a), La prédication verbale et les cadres prédicatifs, Louvain, Peeters.

FRANÇOIS J. (2003b), La *Role and Reference Grammar*, une grammaire de l'interface entre syntaxe, sémantique et pragmatique, *revue LINX*, pp. 77-89.

FRANÇOIS J., MANGUIN J.L. (dir., 2004). Le Dictionnaire Electronique du CRISCO : un mode d'emploi à trois niveaux. *Cahiers du CRISCO*, n°17.

FRANÇOIS J. (2004). Polysémie verbale et cadres participatifs : demander et ses synonymes, http://panini.u-paris10.fr/dea/?u_act=download&dfile=Francois3.pdf.

FRANÇOIS J., VICTORRI B., MANGUIN J.L. (2005). Polysémie adjectivale et synonymie: l'éventail des sens de *curieux*, in SOUTET O. (ed.) *La polysémie*, presses de l'Université de la Sorbonne.

FREROT C., BOURIGAULT D., FABRE C. (2003). Marier procédures d'apprentissage endogènes et ressources exogènes dans un analyseur syntaxique de corpus – Le cas du rattachement verbal à distance de la préposition « *de* » , *TAL*, n°44-3.

FREROT C. (2003). Procédures d'apprentissage endogène doublées de ressources exogènes : résolution en corpus d'une ambiguïté sur « de ». *Actes de la Conférence TALN*, session étudiante *Récital*, Batz-sur-Mer, 11-14 juin, pp. 459-468.

FUCHS C. (1982). *La paraphrase*, Paris, PUF.

FUCHS C., DANLOS L., LACHERET-DUJOUR A., LUZZATI D., VICTORRI B. (1993). *Linguistique et Traitements Automatiques des Langues,* Hachette.

FUNG P., CHURCH K.W. (1994). K-vec : A New Approach for Aligning Parallel Texts. *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto.

GALA, N. (2003). *Un modèle d'analyseur syntaxique robuste fondé sur la modularité et la lexicalisation de ses grammaires*. Thèse de doctorat en informatique. Université de Paris-Sud.

GALA, N. (2003). Une méthode non supervisée d'apprentissage sur le Web pour la résolution d'ambiguïtés structurelles liées au rattachement prépositionnel, *TALN-2003*, Batz-sur-Mer.

GAUME B., DUVIGNAU K., GASQUET O., GINESTE M-D. (2002). Forms of Meaning, Meaning of Forms, *Journal of Experimental and Theoretical Artificial Intelligence*, n°14-1, pp. 61-74.

GAUME B. (2003). Analogie et Proxémie dans les réseaux petits mondes, in DUVIGNAU K., GASQUET O., GAUME B. (eds) Regards croisés sur l'analogie. *Revue d'Intelligence Artificielle*, Vol 5/6. Hermès Sciences.

GIGUET E. (2003). Rapport d'activité CNRS, laboratoire Lattice CNRS UMR 8094 – ENS, Montrouge.

GOLDBERG A. (1995). *Constructions : a construction grammar approach to argument structure*, Chicago and London, University of Chicago Press.

GOVE P.B. (1984). (ed.): *Webster's New Dictionary of Synonyms*. G.&C. Merriam Co.

GREFENSTETTE G (1994). *Explorations in Automatic Thesaurus Discovery*, London, Kluwer Academic Publishers.

GRIMSHAW J. (1991). *Argument Structures*, Cambridge (Mass.), MIT Press.

GROSS G. (1989). Désambiguïsation sémantique à l'aide d'un lexique-grammaire, *Semantica*, Paris, Ladl et Univ. Paris 7.

GROSS G (2004). Réflexions sur le traitement automatique des langues, *Actes de JADT 2004,* Vol. 1 545-556 .

GROSS M. (1975). *Méthodes en syntaxe,* Paris, Hermann.

GROSS M. (1984). Lexicon-Grammar And The Syntactic Analysis Of French, *COLING 1984*, pp. 275-282

GUILLAUME P. (1937). *La psychologie de la forme*, Paris, Flammarion.

HABERT B., NAZARENKO A. (1996). La syntaxe comme marche-pied de l'acquisition des connaissances : bilan critique d'une expérience, *Journées sur l'acquisition des connaissances*, AFIA, Sète.

HABERT B., NAZARENKO A., SALEM, A. (1997). *Les linguistiques de corpus*, Armand Colin, 1997.

HABERT B., FOLCH H. ILLOUZ G. (1999). Sortir des sens uniques : repérer les mots « mouvants » dans le domaine social. *Sémiotiques*, vol. (17). *Dépasser les sens iniques dans l'accès automatisé aux textes*, HABERT B. (resp.), pp. 121-151.

HABERT B., ILLOUZ G., FOLCH H. (2004). Dégrouper les sens: pourquoi, comment?, *Actes de JADT 2004,* Vol. 1, pp. 565-576.

HAYAKAWA S.I. (1994). *Choose the Right Word*. HarperCollins Publishers.

HIRO K., WU H. FURUGORI T. (1996). Word-sense disambiguation with a corpus-based semantic network. *Journal of Quantitative Linguistics*, *3* (3), pp. 244-251.

HIRST G. (1987). Semantic interpretation and the resolution of ambiguity, *Studies in Natural Language Processing*.

HOFMANN T. (1999). Probabilistic Latent Semantic Analysis, *UAI 1999*, pp. 289-296.

IDE N., VÉRONIS J. (1990). Word Sense Disambiguation with very large neural networks extracted from machine readable dictionaries, *Proceedings of the 14th International Conference on Computational Linguistics*.

INKPEN D., HIRST G. (203). Automatic Sense Disambiguation of the Near-Synonyms in a Dictionary Entry. *CICLing 2003*. pp. 258-267.

JACKENDOFF (1990). *Semantic Structures*. Cambridge (Mass.), MIT Press.

JACQUET G. (2002). *La grammaire cognitive au service de la désambiguïsation du sens : expérimentation sur le verbe* jouer, mémoire de DEA de l'université Paris 6.

JACQUET G. (2004). Using the construction grammar model to disambiguate polysemic verbs in French, *Actes de ICCG3 (International Conference on Construction Grammar),* Marseille.

JACQUET G. (soumis). A model of disambiguation of polysemic verbs in French, *11th conference of the EACL*, 3-7 avril 2006.

JACQUET G. VENANT F. (2005). Construction automatique de classes de sélection distributionnelle, *Actes du colloque TALN*, Dourdan.

JACQUET G., VENANT F., VICTORRI B. (2005), Polysémie lexicale, in ENJALBERT P. (Dir.), *Semantique et TALN*, Paris, Hermès.

KAY P. (2001). Argument Structure Constructions and the Argument-Adjunct Distinction, *Actes de ICCG1*, Berkeley, p. 30.

KILGARRIFF A. (1992). *Polysemy,* Unpublished doctoral dissertation, University of Sussex, United Kingdom.

KILGARRIFF A. (1998). SENSEVAL: An exercise in evaluating word sense disambiguation programs. *8th International Congress on Lexicography (EURALEX-1998)*, pp. 176-174.

KILGARRIFF A. ROSENZWEIG J. (2000). English SENSEVAL : Report and results. *2nd International Conference on Language Resources and Evaluation (LREC-2000)*, *3*, pp. 1239-1244.

KILGARRIFF A., GREFENSTETTE G. (2003). Introduction to the Special Issue on the Web as Corpus. Computational Linguistics, n°29-3, pp. 333-348.

KLEIBER G. (1990). *La sémantique du prototype : catégories et sens lexic*al, Paris, P.U.F..

KLEIBER G. (1994). *Nominales : essai de sémantique référentielle*, Paris, Armand Colin.

KLEIBER G. (1999). *Problèmes de sémantique : la polysémie en question*, Paris, presse Universitaire du Septentrion.

LAKOFF G. (1987). *Women, Fire, and Dangerous Things*, Chicago, University of Chicago Press.

LAMBRECHT K. (1994). Information Structure and Sentence Form : a Theory of Topic, Focus, and the Mental Representation of Discourse Referents, Cambridge, Cambridge University Press.

LANGACKER R. W. (1987). *Foundations of Cognitive Grammar, vol. 1 : Theoretical Prerequisites*, Stanford University Press.

LELAND T. (2001). *la polysémie lexicale : l'articulation entre la signification et la référence. Etude comparative de trois polysèmes en français et en anglais*, thèse de doctorat de l'Université de Paris 8.

LESK M. (1986). Automatic sense disambiguation using machine readable dictionaries : How to tell a pine cone from an ice cream cone. *Special Interest Group for DOCumentation (SIGDOC-1986)*, n°17-4, pp. 24-26.

LEVIN B. (1993). *English Verb Classes and Alternations*, Chicago, University of Chicago Press.

LIN D., PANTEL P. (2001). Induction of Semantic Classes from Natural Language Text, *Actes de ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001*.

DE LOUPY C., BELLOT P., EL-BÈZE M., MARTEAU P.F. (1998). Query Expansion and Classification of Retrieved Documents. *TREC 1998*, pp. 382-389.

MANGUIN J.L., VICTORRI B. (1999). Représentation géométrique d'un paradigme lexical, *Actes de la 8ème conférence TALN*, vol. 1, pp. 363-368.

MANGUIN J.L. (2001). Construction d'espaces sémantiques associés aux verbes de déplacement d'objets à partir des données des dictionnaires informatisés des synonymes, *Syntaxe et Sémantique*, n°2, pp. 287-300.

MANGUIN J.L., FRANÇOIS J., VICTORRI B. (2002). « Polysémie adjectivale et rection nominale: quand *gros* et *gras* sont synonymes », in FRANÇOIS J. (ed.), *L'adjectif en français et à travers les langues*, presses Universitaires de Caen, 2002.

MANGUIN J.L., FUJIMURA I., JACQUET G. VENANT F. (à paraître), Etude simultanée de la synonymie et de l'antonymie*, 4èmes journées de linguistique de corpus*, 15-17 septembre 2005, Université de Bretagne Sud, Lorient.

MANNING, C. D. AND SCHÜTZE, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

MCROY S. (1992). Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, n°18, pp. 1-30.

MEYER D., SCHVANEVELDT R. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, n°90-2, pp. 227-234.

MICHAELIS L. A. (2003). Word Meaning, Sentence Meaning and Constructional Meaning. In Cuyckens H., DIRVEN R. ET TAYLOR J. (eds.), *Cognitive Perspectives on Lexical Semantics.* Amsterdam: Mouton de Gruyter, pp. 163-210.

MIHALCEA R., MOLDOVAN D. (1998). Word Sense Disambiguation based on semantic density. *Workshop on Usage of Wordnet Natural Language Processing Systems (COLING-ACL-1998).*

MIHALCEA R., CHKLOVSKI T. (2003). Building sense tagged corpora with volunteer contributions over the Web,. *RANLP 2003*, pp. 357-366.

MULLER P., SARDA L. (1999). Représentation de la sémantique des verbes de déplacement transitifs directs du français, *Revue TAL*, n°39-2, pp. 127-147.

NAZARENKO A. (1998), (Dir.) Compositionnalité, *Revue TAL*, n°39-1.

NAZARENKO A., ZWEIGENBAUM P., HABERT B., BOUAUD J. (2001). Corpus-based Extension of a Terminological Semantic Lexicon. In *Recent Advances in Computational Terminology*, BOURIGAULT D., JACQUEMIN C., AND L'HOMME M.C. (eds), John Benjamins, Amsterdam.

NG H. T., LEE Y. K. (1996). Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. *34thAnnual Meeting of the Society for Computational Linguistics n*°17-4, pp. 40-47.

NG H. T. (1997). Getting serious about word sense disambiguation, *Association for Computational Linguistics Special Interest Group on the Lexicon (ACLSIGLEX-1997) : Workshop* « *Tagging Text with Lexical Semantics : Why, What, and How?* », pp. 1-7.

NG H. T., LEE Y. K. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. *7th Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pp. 41-48.

NUNBERG G. (1978). *The pragmatics of reference*, Indiana University Linguistics Club.

NUNBERG G. (1995). Transfers of Meaning, *Journal of Semantics*, n°17, pp. 109-132.

NUNBERG G., ZAENEN A. (1997). « La polysémie systématique dans la description lexicale », *Langue Française*, 113, 12-23, 1997.

NYCKEES V. (1998). *La sémantique*, Paris, Belin.

PAILLARD D. (2001). À propos des verbes `polysémiques': identité sémantique et variation, *Syntaxe et sémantique 2*, presses universitaires de Caen, pp. 99-120.

PANTEL P., LIN D. (2002). Discovering word senses from text, *KDD 2002*, pp. 613-619.

PANTEL P., RAVICHANDRAN D. (2004). Automatically Labeling Semantic Classes. In *Proceedings of Human Language Technology / North American chapter of the Association for Computational Linguistics (HLT/NAACL-04)*, Boston, MA, pp. 321-328.

PELLETIER J.F. (1975). Non Singular Reference, in PELLETIER J.F. (ed.), *Mass terms: Some Philosophical Problems*, Reidel, Dordrecht, pp. 1-14.

PEREIRA F., TISHBY N., LEE L. (1993). Distributional clustering of English words. *31st Annual Meeting of the Association for Computational Linguistics (ACL-1993)*, pp. 183-190.

PIERREL J.M. (2000). (ed.) *Ingénierie des langues*, Paris, Hermès.

PICOCHE J. (1986). *Structures sémantiques du lexique français*, Nathan.

PLOUX S., VICTORRI B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes, *TAL*, n°39-1, pp. 161-182.

POTTIER B. (1987). *Théorie et analyse en linguistique*, Paris, Hachette.

PUSTEJOVSKY J. (1995). *The generative lexicon*, Cambridge, MIT Press.

PENG Q., ITO T., FURUGORI T. (2001). Word Sense Disambiguation with a Corpus-Based Semantic Network, *NLPRS 2001*, pp. 75-82.

RAPPAPORT HOVAV M., LEVIN B. (1998). Building Verb Meanings, in BUTT M., GEUDER W. (eds.) *The Projection of Arguments: Lexical and Compositional Factors,* CSLI Publications, Stanford, 97-134.

RASTIER F., CAVAZZA M., ABEILLE A. (1994). *Sémantique pour l'analyse – de la linguistique à l'informatique*, Masson.

RESNIK P. (1995). Disambiguating noun groupings with respect to WordNet senses. *3th Workshop on Very Large Corpora*, pp. 54-68.

RESNIK P. (1997). Selectional preference and sense disambiguation, *Association for Computational Linguistics Special Interest Group on the Lexicon (ACL-SIGLEX-1997) : Workshop « Tagging Text with Lexical Semantics : Why, What, and How? »*, pp. 95-130.

RESNIK P. (1999). Mining the Web for Bilingual Text, *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, College Park, Maryland.

RICOEUR P. (1975). *La métaphore vive*, Seuil.

RIVEST R. L. (1987). Learning decision lists, *Machine Learning*, n°2, pp. 229-246.

ROMERO-LOPES M. C. (2002). Identité et variation du verbe *jouer*, *Langue française*, Paris, Larousse.

SALTON G. (1989). *Automatique Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley, Reading.

SARDA L. (1999). *Contribution à l'étude de la sémantique de l'espace et du temps : analyse des verbes de déplacement transitifs directs du français*. Thèse de doctorat (Sciences du langage), Équipe ERSS, Université de Toulouse II.

SCHÜTZE H. (1992). Dimensions of meaning. *Supercomputing-1992*, pp. 787-796.

SCHÜTZE H. (1998). Automatic word sense discrimination. *Computational Linguistics : Special Issue on Word Sense Disambiguation*, n°24-1, pp. 97-123.

SEARLE J. R. (1983). *Intentionality: an Essay in the Philosophy of Mind*, Cambridge, Cambridge University Press.

SEBER G.A.F. (1984). *Multivariate Observations*, Wiley, New York, pp. 317-322.

STEIN A. (1999). Describing Verb Semantics in a Type Hierarchy. Disambiguation of Italian Verbs. In SAINT-DIZIER P. (ed.), *Predicative Forms in Natural Language and in Lexical Knowledge Bases*. Dordrecht: Kluwer, (Text, Speech and Language Technology), pp. 111-137.

STEVENSON M. (1998). Extracting syntactic relations using heuristics. *Proceedings of the European Summer School on Logic, Language and Information'98,* Saarbr/icken, Germany.

STEVENSON M., WILKS Y. (2001). The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, n°27-3, pp. 321-349.

STRAPPARAVA C., GLIOZZO A., GIULIANO C. (2004). Pattern abstraction and term similarity for word sense disambiguation: first at senseval-3, *Proceedings of SENSEVAL-3 Third International Workshop on Evaluation of Systems for the Semantic Analysis of Text*, Barcelone, pp. 229-234.

SUSSNA M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. Proceedings of the Second International Conference on Information and Knowledge Base Management, CIKM'93, Arlington, Virginie, pp. 67-74.

VENANT F. (2002). *Polysémie adjectivale et calcul du sens*, mémoire de DEA de Sciences Cognitives, Paris, EHESS.

VENANT F. (2004). Polysémie et calcul du sens, *Le poids des mots*, *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles* (*JADT 2004*), Vol 2, pp. 1146-1157.

VANDELOISE C. (1993). La préposition *à* pâlit-elle devant *toucher* ?, *Langages*, n°110, p. 107.

VAN VALIN R. (2004). *The Syntax-Semantics-Pragmatics Interface: an introduction to Role and Reference Grammar*, Cambridge, Cambridge University Press.

VERGNE J., GIGUET E. (1998). Regards Théoriques sur le « Tagging ». *Actes de TALN 1998*, Paris, 10-12 Juin.

VERGNE J. (2001). Analyse syntaxique automatique de langues : du combinatoire au calculatoire (communication invitée), *Actes de TALN 2001*, pp. 15-29.

VÉRONIS J. (1998). A study of polysemy judgements and inter-annotator agreement. *Programme and Advanced Papers of the Senseval Workshop*.

VÉRONIS J. (2003). Cartographie lexicale pour la recherche d'information. *10ème conférence sur le Traitement Automatique des Langues Naturelles (TALN-2003)*, pp. 265-274.

VERONIS, J. (2004). Quels dictionnaires pour l'étiquetage sémantique ?, FUCHS ET B. HABERT (resp.), *Le français moderne, Traitement automatique et ressources numérisées pour le français*, n°72-1, pp. 27-38.

VICTORRI B., FUCHS C. (1996). *La polysémie, construction dynamique du sens*, Paris, Hermès.

VICTORRI B. (1997). La polysémie : un artefact de la linguistique ?, *Revue de Sémantique et de Pragmatique*, n°2, pp. 41-62.

VICTORRI B. (1999). Le sens grammatical, *Langages*, Paris, Larousse, p. 136.

VICTORRI B. (2002). Espaces sémantiques et représentation du sens, *Textualités et nouvelles technologies*, éc/artS, 3.

VICTORRI B. (2005). Le calcul de la référence, in ENJALBERT P. (éd.), *Sémantique et TALN*, Paris, Hermès.

VILLE-OMETZ F. (2000). La préposition *à* et les verbes de transfert de disposition, in FRANÇOIS J. (éd.), *Syntaxe et sémantique 2 : sémantique du lexique verbal*, presse Universitaire de Caen, pp. 139-158.

VOLK M. (2001). Exploiting the WWW as a corpus to resolve PP attachment ambiguities. *Proceeding of Corpus Linguistics* 2001. Lancaster.

WEAVER W. (1955). Translation, in WILLIAM N., BOOTH A. D. (Eds.), *Machine translation of languages,* New-York, pp. 15-23.

WILKS Y. (1975). Preference semantics. *Formal Semantics of Natural Language*, pp. 329-348.

WILKS Y., STEVENSON M. (1998). Word Sense Disambiguation using optimised combinations of knowledge sources. *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL-1998)*, pp. 1398-1402.

WITTGENSTEIN L. (1961). *Tractatus logico-philosophicus*, suivi de *Investigations philosophiques* (trad. de l'allemand 1953), Paris, Gallimard.

YAROWSKY D. (1992). Word sense disambiguation using statistical models of roget's categories trained on large corpora. *14th International Conference on Computational Linguistics (COLING-1992)*, pp. 454-460.

YAROWSKY D. (2000). Hierarchical decision list for word sense disambiguation, *Computers and the humanities*, Netherlands, Kluwer Academic Publishers, Vol. 34, pp. 179-186.

ZIPF G. K. (1945). The meaning-frequency relationship of words. *Journal of General Psychology*, n°33, pp. 251-266.

ZWEIGENBAUM P., HABERT, B. (2004). Accès mesurés au sens. *Mots*, vol. 74, ENS Editions, pp. 93-106.