# Ressources et méthodes semi-supervisées pour l'analyse sémantique de texte en français

Claire Mouton

Université Paris 11 - Paris Sud
UFR d'informatique

Thèse en vue de l'obtention du diplôme de
docteur de l'université de Paris 11
en Informatique

# Ressources et méthodes semi-supervisées pour l'analyse sémantique de texte en français

Claire MOUTON

| | | |
|---|---|---|
| Rapporteurs : | M^me^ | Claire GARDENT |
| | M. | Emmanuel MORIN |
| Directrice : | M^me^ | Anne VILNAT |
| Co-directeur : | M. | Gaël DE CHALENDAR |
| Examinateurs : | M. | Gregory GREFENSTETTE |
| | M. | Joseph MARIANI |
| | M. | Benoît SAGOT |

Paris - 16 juin 2011

**Remerciements** Je souhaite ici exprimer ma gratitude et ma joie à toutes les personnes qui m'ont suivies et encouragées durant ces trois dernières années. Que ce soit dans la sphère professionnelle ou dans la vie privée, d'un simple mot d'encouragement à un suivi scientifique régulier, à tous un très grand merci, car c'est grâce à vous que j'en suis aujourd'hui à rédiger ces lignes, signe d'un aboutissement certain.

Parmi cette non-liste anonyme et néanmoins chargée de sens figure un certain nombre de *Happy Few*. La délation n'étant qu'un crime mineur en comparaison de l'emploi de ce si joyeux anglicisme dans une thèse sur le traitement de la langue française, je prends donc la liberté de dénoncer ces personnes et de les faire par là même sortir de l'anonymat et entrer dans ladite catégorie.

Je souhaite donc en premier lieu exprimer ma reconnaissance à M. Gaël de Chalendar, mon encadrant le plus direct. Tout d'abord pour m'avoir accordé sa confiance en me proposant ce sujet de thèse il y a trois ans. Puis, durant ces trois années elles-mêmes, pour ses conseils scientifiques et techniques ainsi que pour la constance de ses encouragements et de sa sympathie. Ils constituent la recette d'un encadrement réussi et m'ont permis de trouver les ressources personnelles nécessaires pour mener ces travaux à bien.

Je souhaite tout autant remercier Mme Anne Vilnat, pour avoir accepté de reprendre la direction de ma thèse en cours de route, en supplément de ses autres nombreuses responsabilités. Et surtout de l'avoir fait avec un regard aussi consciencieux, scientifique et chaleureux, malgré la distance qui séparait nos sites respectifs de travail.

Mes remerciements vont également à M. Gregory Grefenstette, pour son encadrement complémentaire, pour ses conseils ponctuels et avisés, pour ses pointeurs bibliographiques ainsi que pour les diverses inspirations qu'il a su insuffler à mon travail.

Je fais part de toute ma gratitude à M. Joseph Mariani qui a accepté de présider le jury de soutenance, ainsi qu'à mes rapporteurs et examinateur Mme Claire Gardent, M. Emmanuel Morin et M. Benoît Sagot.

Enfin, tout ceci n'aurait pas été possible sans l'accueil chaleureux des équipes du CEA List, des labs d'Exalead ainsi que du Limsi. Une pensée toute particulière va à Julien C. pour sa confiance et son accompagnement lors des premières années de ce projet, à Guillaume pour m'avoir mis le pied à l'étrier ainsi qu'à Pierre-Alain pour toutes ses leçons de vie dont la qualité a su rester constante jusqu'à aujourd'hui. Je remercie également Myriam Rakho, Benoît Richert et Adrien Walkowiak pour leur implication et l'émulation qui a résulté des travaux que nous avons menés en commun. Merci de leur sympathie et de leur soutien à tous ceux que je n'ai pas cité mais que j'ai côtoyé avec plaisir durant ces trois années.

Enfin, merci à tous mes proches et amis qui m'ont vu endurer les affres de la thèse et parvenir avec bonheur au terme de cette aventure.

**Résumé** Pouvoir chercher des informations sur un niveau sémantique plutôt que purement lexical devrait améliorer la recherche d'informations. Cette thèse a pour objectif de développer des modules d'analyse sémantique lexicale afin d'améliorer le système de recherche de documents textuels de la société Exalead. Les travaux présentés concernent plus spécifiquement l'analyse sémantique de texte en français. La problématique liée au traitement du français réside dans le fait qu'il n'existe que peu de ressources sémantiques et de corpus annotés pour cette langue. Rendre possible une telle analyse implique donc d'une part de pourvoir aux besoins en ressources linguistiques françaises, et d'autre part, de trouver des méthodes alternatives ne nécessitant pas de corpus français manuellement annoté. Notre manuscrit est structuré en trois parties suivies d'une conclusion.

Les deux chapitres de la première partie délimitent les objectifs et le contexte de notre travail. Le premier introduit notre thèse en évoquant la problématique de la sémantique en recherche d'information, en présentant la notion de sens et en identifiant deux tâches d'analyse sémantique : la *désambiguïsation lexicale* et l'*analyse en rôles sémantiques*. Ces deux tâches font l'objet de l'ensemble de notre étude et constituent respectivement les parties 2 et 3. Le second chapitre dresse un état de l'art de toutes les thématiques abordées dans notre travail.

La deuxième partie aborde le problème de la désambiguïsation lexicale. Le chapitre 3 est consacré à la constitution de nouvelles ressources françaises pour cette tâche. Nous décrivons dans un premier temps une méthode de traduction automatique des *synsets* nominaux de WordNet vers le français à partir de dictionnaires bilingues et d'espaces distributionnels. Puis, nous constituons une ressource automatiquement en proposant une adaptation de deux méthodes d'induction de sens existantes. L'originalité des clusters de sens ainsi constitués est de contenir des mots dont la syntaxe est proche de celle des mots source. Ces clusters sont alors exploités dans l'algorithme que nous proposons au chapitre 4 pour la désambiguïsation elle-même. Le chapitre 4 fournit également des recommandations concernant l'intégration d'un tel module dans un système de recherche de documents.

L'annotation en rôles sémantiques est traitée dans la troisième partie. Suivant une structure similaire, un premier chapitre traite de la constitution de ressources pour le français, tandis que le chapitre suivant présente l'algorithme développé pour l'annotation elle-même. Ainsi, le chapitre 5 décrit nos méthodes de traduction et d'enrichissement des prédicats de FrameNet, ainsi que l'évaluation associée. Nous proposons au chapitre 6 une méthode semi-supervisée exploitant les espaces distributionnels pour l'annotation en rôles sémantiques. Nous concluons ce chapitre par une réflexion sur l'usage des rôles sémantiques en recherche d'information et plus particulièrement dans le cadre des systèmes de réponses à des questions posées en langage naturel.

La conclusion de notre mémoire résume nos contributions en soulignant le fait que chaque partie de notre travail exploite les espaces distributionnels syntaxiques et que ceci permet d'obtenir des résultats intéressants. Cette conclusion mentionne également les perspectives principales que nous inspirent ces travaux. La perspective principale et la plus immédiate est l'intégration de ces modules d'analyse sémantique dans des prototypes de recherche documentaire.

**Abstract** The possibility of performing semantic rather than purely lexical search should improve information retrieval. This Ph.D work aims at developing modules of lexical semantic analysis, having as a further objective to improve the textual search engine of Exalead company. Presented works deal more specifically with semantic analysis on the French language. Processing of French language is more complex due to the lack of semantic resources and corpora for this language. Thus, make such an analysis possible implies on the one hand to provide for needs of French linguistic resources, and on the other hand, to find alternate methods which do not require any manually annotated French corpus. Our thesis is divided in three main parts followed by a conclusion.

The first part is composed of two chapters which define the objectives and the context of our work. The first of them introduces our thesis. It evokes some semantic issues in the field of Information Retrieval, then tries to define the notion of sense. Finally, it identifies two semantic analysis tasks, namely *word sense disambiguation* and *semantic role labeling*. These two tasks are the two main topics we address in our whole study. They are respectively handled in part 2 and 3. The second chapter draws up a state-of-the-art review of all the topics addressed in our work.

The second part tackles the word sense disambiguation issue. Chapter 3 is devoted to the building of new French resources dedicated to this task. We first describe a method to automatically translate the nominal *synsets* of WordNet to French, by using bilingual dictionaries and distributional spaces. Secondly, we put forward an adaptation of two existing methods of word sense induction, in order to acquire a word senses resource in a fully automatic way. Moreover, the sense clusters built in the latter step show originality as they contain words whose syntax is similar to the the syntax of the given ambiguous words. The so-called sense clusters are then used in the word sense disambiguation algorithm that we put forward in chapter 4. This chapter also provides recommendations in order to integrate such a module in a textual search engine.

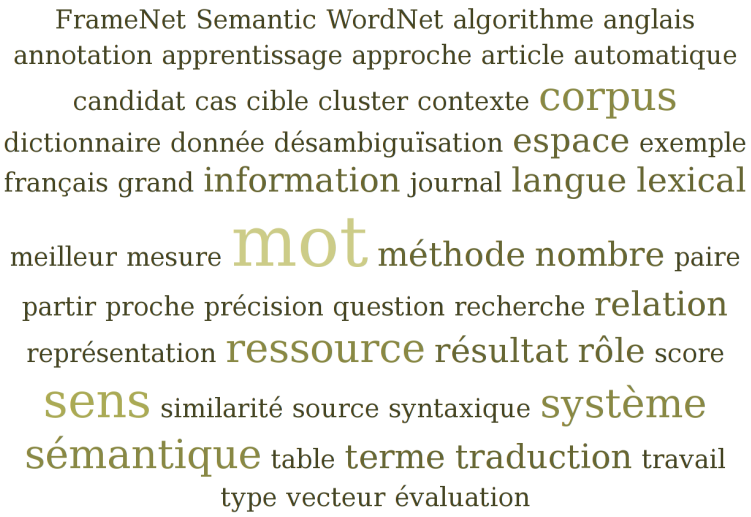
Semantic role labeling is handled in the third part. In a similar fashion, a first chapter deals with the building of resources for the French language, whereas the following chapter presents the algorithm developed for the labeling task itself. Chapter 5 thus describes the method we propose to translate and enrich FrameNet predicates, as well as the related evaluation. We propose in chapter 6 a semi-supervised approach which uses the distributional spaces to label semantic roles. We conclude this chapter with some considerations on the use of semantic roles in information retrieval and more specifically in the scope of question answering systems.

The conclusion of our thesis summarizes our contributions. It emphasizes the fact that each step of our work uses syntactical distributional spaces and that it provides interesting results. This conclusion also draws the main perspectives we see to pursue our studies. The main and immediate concern is to integrate these semantic analysis modules into prototypes for textual documents search.

# Table des matières

# Première partie

# Objectifs et contexte


FrameNet Semantic WordNet algorithme anglais
annotation apprentissage approche article automatique
candidat cas cible cluster contexte corpus
dictionnaire donnée désambiguïsation espace exemple
français grand information journal langue lexical
meilleur mesure mot méthode nombre paire
partir proche précision question recherche relation
représentation ressource résultat rôle score
sens similarité source syntaxique système
sémantique table terme traduction travail
type vecteur évaluation

# Chapitre 1

# Introduction

## 1.1 Des traitements sémantiques pour la Recherche d'Information

Dans le cadre de la recherche de documents textuels, les moteurs de recherche sont entre autres confrontés à deux problématiques linguistiques issus des caractéristiques sémantiques du vocabulaire.

La première problématique évoquée provient du phénomène de synonymie. Un cas d'école pour illustrer ce problème est la requête comportant les trois mots-clés suivants : *école, port, voile*. L'utilisateur a une intention bien précise, néanmoins le moteur ne sait pas s'il fait référence à *l'école de voile du port* ou au *port du voile à l'école*.

Les moteurs de recherche classiques ne traitent pas les mots grammaticaux et traiteraient ces trois requêtes de la même façon. En revanche, une analyse syntaxique et sémantique des deux requêtes comportant les mots grammaticaux permettrait de les désambiguïser et de ne retourner que les résultats correspondant aux intentions de l'utilisateur.

Dans un cas ou une telle désambiguïsation n'est pas possible (requête sans mots grammaticaux par exemple), alors il faudrait que le moteur soit malgré tout capable de distinguer les deux sens afin de présenter les documents concernant chaque sujet de façon distincte.

Outre le phénomène linguistique de polysémie que nous venons de voir, la recherche d'information souffre également du phénomène de synonymie et même de similarité sémantique. En effet, toujours en référence à l'exemple précédent, imaginons deux documents dont l'un contiendrait les mots *club nautique de la marina* et l'autre *port de la burqa dans les lieux d'enseignement*.

Bien que l'intégralité des termes de la requête ne soit présente dans aucun de ces deux documents, l'utilisateur sera néanmoins intéressé par ces deux documents (respectivement à la requête d'origine).

Les problèmes d'analyse du sens sont également soulevées dans le cadre des systèmes de réponses à des questions posées en langage naturel. En effet, il est très simple pour l'humain d'établir la correspondance entre la question *Quelle est la route qu'il faut prendre pour se rendre de Lyon à Marseille ?* et la phrase *L'autoroute du Soleil est une autoroute française qui prolonge l'autoroute A6 au niveau de Lyon et qui va jusqu'à Marseille.* pour déterminer que la réponse est *l'autoroute du Soleil.* Cependant, cela implique des traitement sémantiques et logiques complexes qui ne sont pas encore totalement résolus à l'heure actuelle par les systèmes automatiques.

Dans cette introduction, nous tentons tout d'abord de définir les limites de la notion de sens. Puis nous procédons à l'identification des besoins généraux en analyse sémantique. Nous présentons ensuite deux tâches d'analyse sémantique que nous pensons utiles pour la résolution des problèmes de recherche d'information évoqués ci-dessus. C'est sur la résolution de ces tâches en langue française que portent l'ensemble de nos travaux et nous terminons cette introduction par le plan détaillé des différentes étapes auxquelles nous avons procédé.

## 1.2 Mots, sens et ressources linguistiques

Chaque mot de sa langue maternelle évoque quelque chose à l'être humain. Cette représentation fait appel à des images mentales et/ou à une reformulation par d'autres mots. Elle est le résultat d'un long processus d'accords passés entre les hommes. Ce consensus débute lors de l'émergence du langage et est modifié tout au long du temps par les contextes historiques et culturels qui le modèlent encore aujourd'hui. En conséquence, chaque mot réfère à une approximation, à une représentation que les hommes se font des concepts qui les entourent et sur lesquels ils ont apposé des noms qui leur sont communs. Chacun percevant le monde qui l'entoure à sa façon, les mêmes mots sont mis sur des représentations quelque peu différentes des mêmes concepts. Ainsi l'évocation d'une *chaise* appelle des images diverses chez chacun d'entre nous. Il en va de même pour les mots abstraits, mais ceci à une échelle encore plus grande. Par exemple, l'évocation de la *liberté* ou de la *paix* appelle un contenu sémantique dépendant encore plus fortement des connaissances de tout un chacun. L'illustration que nous nous faisons des mots du langage dépend profondément de notre vécu et du contexte dans lequel l'évocation est faite.

Afin de converger au plus proche possible d'une représentation commune, les linguistes se sont intéressés à la définition de sens et à la constitution de dictionnaires. Les premiers dictionnaires connus datent du $IV^e$ millénaire avant J.C. Ces simples listes lexicales sont trouvées chez les Sumériens qui cultivent une très grande tradition lexicographique et répertorient tout le vocabulaire de leurs dialectes ([Bejoint & Thoiron 1996]). Pour la langue française, le premier dictionnaire est constitué sous François $1^{er}$, peu de temps après que la langue française devint la langue administrative. En 1539, le *Dictionnaire françois-latin contenant les motz et manieres de parler françois tournez*

*en latin*[1] de Robert Estienne voit le jour. Il contient les mots français, suivis de leurs traductions latines et d'exemples, ainsi que d'une explication quand le terme est difficile.

Aujourd'hui les dictionnaires répertorient les différents sens que l'on peut distinguer pour chaque mot. Cependant, cette distinction est faite dans l'hypothèse que la distribution de ces sens est discrète et dénombrable. Or, on peut aussi considérer que certains de ces différents sens sont plutôt différents usages du même concept ([Kilgarriff 1997]). Plusieurs phénomènes liés sont observables parmi lesquels nous citerons la métaphore, l'auto-hyponymie, et la lexicalisation de nouveaux concepts.

Rappelons simplement pour la métaphore qu'elle consiste en une comparaison implicite avec un élément (souvent également implicite) du discours, comme dans *Il pleut des cordes* où les *gouttes de pluie* sont un élément implicite du discours implicitement comparées à des *cordes*.

L'auto-hyponymie est la caractéristique qu'un mot peut avoir à être une généralisation ou une spécification de lui-même. Ainsi, le statut du mot *knife* est discuté par [Cruse 1995]. Nous rapporterons ici la discussion sur le mot français *couteau* (référant à l'objet tranchant et non au coquillage). Employé seul, il peut référer à la fois à *un objet tranchant* de manière générale (nous noterons *couteau#1* dans ce cas-là) mais aussi à un *couteau de table* (*couteau#2*) ou à un *couteau à cran d'arrêt* (*couteau#3*). Il est tout à fait plausible qu'une personne déclare *ne pas avoir de couteau* en étant à table même si un canif est posé sur la table, ou inversement et de façon moins réaliste, qu'un bandit annonce *ne pas avoir de couteau* alors qu'il possède des couteaux de table chez lui. Les deux derniers usages (*couteau#2* et *couteau#3*) appartiennent à la catégorie des *couteaux#1* : on dit qu'ils sont hyponymes de *couteau#1* et même auto-hyponymes car le mot *couteau* est alors hyponyme du mot *couteau*, soit lui-même.

Enfin la lexicalisation d'un concept s'effectue lorsque qu'un concept devient trop courant pour que l'on y réfère raisonnablement par une paraphrase. Par exemple, la popularisation de l'utilisation du *système hypertexte public fonctionnant sur Internet qui permet de consulter, avec un navigateur, des pages accessibles sur des sites*[2] a rendu nécessaire sa lexicalisation par l'intermédiaire du mot *Web* en anglais ou *toile* en français. On remarque que la lexicalisation de nouveaux concepts s'inspire souvent de métaphores (comme ici avec le réseau de la toile d'araignée) ou d'autres figures de style où le glissement sémantique laisse une trace.

Les différents usages d'un mot varient en fréquence selon le type de texte que l'on rencontre, ils peuvent à l'extrême, ne pas apparaître dans un genre donné et être essentiels dans un autre. C'est le cas de l'exemple donné par Kilgarriff : le mot *handbag*[3], répertorié comme n'ayant qu'un seul sens dans le LDOCE3, réfère de façon ponctuelle à d'autres concepts. Dans la presse spécialisée musique il s'agit d'un style musical des années 90, le nom ayant été donné par référence aux groupes de filles dansant en cercle autour de leur *handbag* sur cette musique. Le tout venant n'a pas besoin

1. `http://gallica.bnf.fr/ark:/12148/bpt6k505878`
2. Définition donnée par Wikipédia `http://fr.wikipedia.org/wiki/World_Wide_Web`
3. en français : *sac à main*

d'une telle connaissance pour interpréter correctement son quotidien ; en revanche une application de traitement automatique des langues sur le domaine de la musique qui disposera de cette information n'en sera que plus robuste.

Deux des limitations principales rencontrées par les lexicographes traditionnels sont la limite de temps et d'espace : ils ne peuvent ni prendre le temps de répertorier tous les usages existants de chacun des mots du vocabulaire, ni prendre la place nécessaire à un tel répertoire pour l'édition d'un dictionnaire papier. C'est pourquoi la plupart des dictionnaires ne répertorient que les usages les plus courants d'un mot, ou bien ne distinguent pas leur fréquence d'usage, ou encore ne donnent pour définition que le plus petit dénominateur commun à plusieurs usages, sans fournir les caractéristiques qui permettraient leur distinction (comme par exemple la définition de *couteau#1*).

Une approche alternative aux dictionnaires de sens tels qu'on les connaît est la constitution de groupes d'exemples d'usage qui permettrait alors à un système (ou à un humain) de déterminer si l'instance d'un mot ambigu dans une nouvelle phrase est semblable à un groupe existant ou à un autre. C'est à partir de cette idée que certains comme [Pantel & Lin 2002], [Véronis 2004], [Ferret 2004], analysent automatiquement de grands corpus afin de constituer pour chaque mot du vocabulaire, des clusters de mots apparaissant souvent dans un même contexte, sans interférer avec les contextes d'un autre cluster. Ils tentent par cette approche de reproduire ce qu'un lexicographe ferait manuellement en distinguant différents exemples d'usage.

Les dictionnaires et autres répertoires de sens ainsi définis sont certes intéressants en eux-mêmes, mais ils trouvent tout particulièrement leur intérêt lorsque l'on se retrouve confronté à un mot que l'on ne comprend pas dans une phrase. L'humain consulte alors son dictionnaire en espérant y trouver une définition qui correspondra à l'usage du mot dans la phrase, ceci afin de pouvoir interpréter correctement l'ensemble de la phrase.

Dans le cadre du traitement automatique des langues, différentes applications vont chercher à extraire "intelligemment" de l'information en fonction de différents besoins définis par leurs concepteurs et utilisateurs. De façon similaire à l'humain, ces systèmes ont besoin d'interpréter l'ensemble d'une phrase. Voyons maintenant en quoi cela peut consister pour de tels systèmes.

## 1.3 Analyse sémantique : pourquoi et comment ?

L'analyse sémantique de texte consiste à analyser de façon automatique le sens des mots. Cette section a pour but de présenter les besoins en analyse sémantique et de définir le cadre de notre étude.

### 1.3.1 Appeler un chat un chat

Malgré les consensus communs sur le sens des mots, bâtis par l'intermédiaire des dictionnaires, les appellations que donnent les hommes à leurs différentes actions et aux différents objets de leur environnement varient énormément d'une personne à l'autre. En effet, [Furnas *et al.* 1987] étudient la variation de l'appellation de différentes actions et objets de domaines variés, par un panel de personnes composé à la fois d'experts et d'utilisateurs tout venant. Leurs résultats montrent que sur les cinq collections d'objets étudiés, deux personnes n'ont en moyenne que moins de 1 chance sur 5 de nommer un même objet par un même mot. Cette probabilité double lorsque l'on choisit pour nom de référence le nom attribué par la majorité des sujets, ou que l'on s'autorise trois appellations de référence prises au hasard parmi les différentes appellations données par les sujets. Les auteurs prônent l'usage de systèmes prenant en compte tout le vocabulaire possible et usant de différentes stratégies pour déterminer à quel objet il est effectivement fait référence. Le fait de proposer tout le vocabulaire en tant qu'accès aux données fait apparaître un nouveau problème : une même appellation peut référer à plusieurs entités différentes, il faut alors lever les ambiguïtés lorsque cela est nécessaire.

Cette étude de cas général met en valeur l'importance de l'analyse et de l'interprétation des requêtes envoyées à un système de recherche d'information. En effet, le premier problème soulevé concerne le lien sémantique des mots employés dans la requête avec l'ensemble des termes susceptibles d'être pertinent à la recherche de l'utilisateur (synonymes, hyponymes, ...). On peut alors exprimer le besoin de mise en correspondance des termes utilisés avec les termes intéressants que ce soit par le biais d'une extension de la requête ou d'une indexation plus complexe que la simple indexation des mots appartenant aux documents.

L'identification du problème d'ambiguïté est également applicable à la recherche d'information, un même mot peut faire référence à des objets différents ou même porter des sens différents. Nous ne mentionnons ici que brièvement ce phénomène dont nous aurons l'occasion de reparler plus en détail dans la suite de ce manuscrit.

### 1.3.2 Différentes visions du sens et de la cognition

Lorsque l'on s'intéresse aux approches cognitives de la représentation du sens et de la modélisation du cerveau, on trouve trois grands courants de pensée ([Harnad 1990]). L'école symbolique propose une représentation où les systèmes de cognition sont structurés à partir de règles explicites dépendant uniquement de symboles et de leur formes. Dans ce paradigme, les symboles du langage naturel sont les mots, et seule leur forme et leurs agencements sont significatifs. Pour simplifier, un dictionnaire monolingue doté de définitions suffirait à décrire l'ensemble du sens de la langue.

L'école connexionniste considère que la cognition n'est pas effectuée par de la manipulation de symboles mais par des réseaux dynamiques d'unités de traitement inter-connectées qui sont contraintes et ajustées en fonction des données d'entrée et des sorties souhaitées. Les algorithmes de réseaux de neurones sont issus de ce modèle de cognition et s'avèrent effectivement efficace pour des traitements de reconnaissance ou d'identification d'objets. Il s'agit d'une modélisation sensorielle (visuelle, sonore ou autre), avec laquelle on peut reconnaître les caractéristiques communes de différents objets et donner une appellation à l'ensemble des objets considérés appartenant à une même catégorie.

D'après Harnad, les deux précédentes théories ne sont pas de bons modèles de cognition. Le courant symbolique fait face au problème de *symbol grounding* : aucun système de raisonnement (machine ou humain) ne peut répondre au test de Turing[4] s'il n'a pour données d'entrée qu'un dictionnaire monolingue et aucune connaissance préalable d'un autre langage. En effet, de définition en définition, le raisonnement tournera en boucle sans jamais comprendre à quelle réalité il est référé, et le système sera incapable de tenir une conversation sensée.

Par ailleurs, un modèle purement connexionniste ne pourra effectuer de raisonnement entre les différentes catégories qu'il est capable d'ancrer dans le monde réel. Pour Harnad, la seule hypothèse possible passe par un modèle ascendant. Dans un tel modèle, pour un certain nombre de symboles élémentaires (terminaux), il existe une représentation iconique définie comme une empreinte sensorielle totale issue de la perception (i.e. pourquoi pas du modèle connexionniste) ainsi que des représentations catégoriques associées, définie comme étant des empreintes issues de l'empreinte totale et ne conservant que les caractéristiques propres à catégoriser la perception donnée. Pour donner un exemple, la représentation iconique de la vision d'un *Saint-bernard* est l'image mentale de ce Saint-bernard-là en action dans son contexte. Tandis que les représentations catégoriques associées seraient la partie nécessaire et suffisante de cette information pour savoir qu'il s'agit d'un Saint-bernard, d'un chien, d'un animal, d'un être vivant...

Parallèlement à ces représentations, il existerait dans ce modèle ascendant une représentation symbolique permettant d'interpréter des symboles complexes (non-terminaux) à partir des représentations iconiques et catégoriques existantes. Par exemple, un modèle ayant connaissance des représentations catégoriques de *cheval* et de *rayure* et n'ayant jamais rencontré de *zèbre* pourrait néanmoins interpréter la règle symbolique *zèbre* = *cheval* + *rayure*.

À notre sens, le problème d'appellation des objets soulevé par [Furnas *et al.* 1987] et décrit précédemment relève à la fois d'un problème de définition de règles symboliques et d'un problème

4. Test dans lequel une intelligence artificielle doit être confrontée avec un testeur humain dans un protocole de dialogue. Si le testeur humain n'arrive pas à déterminer (en un temps infini) s'il s'agit d'un autre humain ou d'une machine, l'intelligence artificielle a passé le test de Turing.

d'ancrage à la réalité. Les objets à nommer n'étant pas nécessairement des unités élémentaires, une partie du problème relève donc de la définition des règles symboliques décrivant notre vocabulaire commun.

L'analyse sémantique au sens où nous l'entendons dans cette thèse sera restreinte à la modélisation symbolique du sens. Aucune considération sur l'ancrage des sens dans la réalité ne sera considérée comme appartenant au cadre de notre recherche.

## 1.4 Deux tâches d'analyse sémantique lexicale

Lors de l'interprétation d'un texte, plusieurs niveaux de connaissance se superposent. L'analyse lexico-syntaxique consiste à découper les mots, syntagmes et phrases, à analyser les genres, nombres, cas grammaticaux, voies actives, passives ou moyennes.., et à déterminer quelles sont les dépendances grammaticales qui lient les mots à l'intérieur d'une phrase. Cette étape est essentielle pour procéder à tout traitement ultérieur, mais certains traitements subtils dépendent des traitements ultérieurs eux-mêmes. Cette interdépendance concerne des cas comme par exemple le rattachement des syntagmes prépositionnels à un syntagme verbal ou à un syntagme nominal. Nous voyons dans les deux phrases *Elsa a acheté à sa fille un livre pour apprendre à lire.* et *Elsa a acheté à sa fille un livre pour l'occuper durant le trajet.* des cas exemples pour lesquels une analyse sémantique serait nécessaire pour déterminer que le syntagme *pour apprendre à lire* se rattache au syntagme *un livre* tandis que le syntagme *pour l'occuper durant le trajet* se rattache à *acheter*.

A partir des informations lexico-syntaxiques, on peut procéder à l'analyse du sens du texte proprement dite. Celle-ci est elle-même divisible en différents niveaux. A gros traits, le niveau le plus bas de l'interprétation du sens correspond à l'analyse du sens des mots et de la composition de ces mots entre eux. On parle alors d'analyse sémantique lexicale.

Quant au niveau le plus haut, il fait appel à des connaissances pragmatiques (connaissances du monde : sens commun ou connaissances encyclopédiques) pour analyser des situations complètes ou des événements, eux-mêmes éventuellement composés de sous-situations (respectivement sous-événements). Il s'agit alors d'une analyse thématique ([Ferret 1998]).

Notre travail se concentre sur les deux principales problématiques en analyse sémantique lexicale. La première concerne l'identification du sens des mots de façon individuelle ou parfois d'une locution figée. La seconde met en jeu l'interaction de ces différents éléments pour former un tout qui donnera lieu à interprétation.

Prenons comme exemple la phrase suivante extraite d'un article de journal :

La table ronde «radio-fréquences, santé, environnement» s'est achevée ce lundi avec une dizaine de pistes mais sans aucune mesure forte.

Cette phrase nous servira d'exemple pour illustrer les deux tâches que nous décrivons ci-dessous.

### 1.4.1 La désambiguïsation lexicale

Pour illustrer le problème de l'ambiguïté lexicale, prenons chacun des mots pleins de notre phrase d'exemple de façon indépendante et observons les différents sens que nous pouvons attribuer à chacun d'entre eux hors contexte. Le tableau 1.1 liste un certain nombre de sens que nous avons pu leur trouver, l'objectif n'étant pas d'être exhaustif sur le sujet. Le lecteur pourra certainement trouver d'autres usages que nous n'aurons pas mentionnés.

| | |
|---|---|
| **table** | meuble<br>tableau (table périodique)<br>cuisine/restaurant (la table d'un grand chef)<br>réunion/conférence (table ronde) |
| **rond** | circulaire/sphérique/cylindrique<br>légèrement gros (pour une personne)<br>saoul<br>sans décimale (pour un chiffre) |
| **santé** | état de l'organisme<br>la prison de la Santé |
| **environnement** | écologie<br>milieu<br>espace virtuel (informatique)<br>voisinage |
| **achever** | terminer<br>tuer |
| **piste** | chemin (piste de 4x4, de ski ou de circuit automobile)<br>idée<br>indice (police par exemple)<br>piste audio |
| **mesure** | dose/portion<br>ratio/indice/score<br>décision de réglementation<br>découpage musical |
| **fort** | puissant/doué<br>gros<br>important<br>incroyable |

TABLE 1.1 – *Dans quel sens lit-on ?*

Nous avons distingué dans ce tableau des sens différents pour chacun de ces mots. On parle d'homonymie lorsque la racine sémantique (et parfois étymologique) des différents sens est réellement différente et que la forme commune de ces mots est le fruit d'un rapprochement ultérieur, aléatoire ou non. C'est le cas du mot *régime* pour lequel le *régime politique* et le *régime amincissant* viennent tout deux du latin *regimen* signifiant *direction, gouvernement*, substantif du verbe d'action *regere*, tandis que le *régime de dattes ou de bananes* est l'adaptation de l'espagnol *racimo* signifiant *grappe de raisin*, et par analogie, *inflorescence de certaines plantes telles que le bananier*. Ce mot espagnol

vient du latin classique *racemus*, *grappe de raisin* qui est aussi l'étymon du mot français *raisin*[5].

En revanche, lorsque la racine sémantique et étymologique du mot est la même et que ce sont les différents usages du mot, parfois familiers ou stylistiques, qui se déclinent en plusieurs variations, on parle alors de polysémie. C'est notamment le cas de la *piste* au sens *piste de ski* qui est dérivée de la *piste* en tant que *chemin* ou de l'*environnement* pour lesquels les différents sens mentionnés réfèrent tous à ce qui environne, ce qui entoure un objet ou une personne.

On distingue encore différents usages d'un même sens bien qu'il soit parfois difficile de déterminer s'il s'agit d'un sens ou d'un usage. Pour exemple, le Wiktionnaire[6] répertorie quatre sens pour le mot *santé* dont trois sont très proches : le bon état de l'organisme, le bon état du moral, et l'état de l'organisme en général qu'il soit bon ou mauvais. Cette nuance entre sens et usage soulève le problème de la granularité de distinction des sens.

Enfin, on remarque aussi des expressions figées comme *table ronde* (*rond* n'a ici aucun sens sans la présence de *table*), ainsi que des expressions métaphoriques ou issues d'autres figures de style comme c'est le cas lorsque l'on emploie *la Santé* pour signifier la prison de la *Santé*.

La tâche de désambiguïsation lexicale consiste à lever l'ambiguïté inhérente à chacun de ces cas. Par abus de langage, les termes *polysémie* ou *polysémique* utilisés dans la suite de ce manuscrit réfèreront sans distinction aux cas d'homonymie, de polysémie, de mots à plusieurs usages, ou aux expressions figées.

Nous présentons dans ces travaux les différentes expérimentations que nous avons menées sur la désambiguïsation lexicale automatique.

### 1.4.2 L'annotation en rôles sémantiques

Nous présentons maintenant le second volet de notre problématique, l'annotation en rôles sémantiques. Le concept de rôle sémantique (autrement appelé rôle thématique ou thêta-rôle, bien que ces deux derniers soient plus spécifiques à des prédicats de type verbal) est introduit très tôt dans l'histoire de la linguistique, puisque la grammaire de la langue indienne de Panini (400 av. J.C) définissait déjà la notion de *karaka* pour exprimer le type de relation liant un argument en relation syntaxique avec un prédicat. Les 6 *karakas* ainsi définis étaient les suivants : *source, bénéficiaire, moyen, lieu, patient, agent* ([Nath Jha 2004]). Les propriétés des rôles sémantiques sont revues et discutées dans la littérature plus récente : [Gruber 1965], [Davidson 1967], [Fillmore 1968], [Jackendoff 1972], [Dowty 1991]. Les divergences de consensus concernent principalement le nombre de ces rôles, la correspondance stricte entre relation syntaxique et rôle thématique (cadre ou grille de sous-catégorisation), ainsi que l'unicité de rôle pour un argument donné (exemple : *Jean courut jusqu'à la maison.* Pour une telle phrase, Gruber et Jackendoff avancent le fait que l'argument *Jean* est à la fois *Agent* et *Thème* transporté) et la distinction des rôles, à savoir, est-ce que deux arguments différents

5. http://projetbabel.org/mots/index.php?p=roi
6. cf. Section 2.2.1.2

peuvent remplir le même rôle ? (exemple : *Jean rencontre Marie*). Nous ne rentrerons pas dans un détail plus avancé de ces propriétés. Retenons simplement la définition de [Saint-Dizier 2006] :

"Un rôle thématique est une étiquette abstraite (par exemple : agent, thème, instrument) qui caractérise la relation sémantique qu'un prédicat (verbe, mais aussi préposition, adjectif, nom prédicatif) peut entretenir avec l'un de ses arguments. "

Si l'on considère l'exemple que nous avons vu précédemment et le référentiel FrameNet [Baker *et al.* 1998] (que l'on présentera de façon plus complète dans le courant de cet exposé), l'annotation en rôles sémantiques donnerait la chose suivante :

| [Activity] | **Activity_finish** | [Time] | [Result] |
|---|---|---|---|
| La table ronde «radio-fréquences, santé, environnement» | **s'est achevée** | ce lundi | avec une dizaine de pistes mais sans aucune mesure forte. |

Dans ces travaux nous adressons le problème de l'annotation automatique de tels constituants.

## 1.5 Structure du manuscrit

Cet exposé se divise en 4 parties distinctes. La première est composée de la présente introduction suivie d'une revue des travaux les plus fondamentaux concernant les différentes thématiques abordées dans ce manuscrit, ainsi que ceux les plus liés à nos propres recherches. La partie II traite du problème de la désambiguïsation lexicale tandis que la partie III s'intéresse à la question de l'annotation sémantique de rôles. La structure de ces deux parties centrales est analogue. Un premier chapitre (3 resp. 5) rassemble les algorithmes que nous proposons et les évaluations que nous avons menées en vue de la génération de nouvelles ressources linguistiques ou de l'enrichissement de ressources existantes. Nous présentons dans le chapitre suivant (4 resp. 6) les expérimentations mises en place pour la tâche d'analyse elle-même et son intérêt en recherche d'information ainsi que les logiciels implémentés dans ce cadre. Enfin, la quatrième partie conclut ce manuscrit en rappelant nos contributions, en discutant les points restant à améliorer, et en proposant des pistes pour de futurs travaux.

# Chapitre 2

# État de l'art

Ayant pour intention de fournir un inventaire détaillé de l'état de l'art des différentes thématiques abordées dans ce manuscrit, nous étudions dans ce chapitre la littérature concernant les différentes problématiques évoquées en introduction.

Les traitements que nous effectuons dans nos travaux faisant appel à des notions avancées de fouille de données, nous dédions la première section de ce chapitre à la présentation de ces notions. Nous présentons dans un premier temps des algorithmes de clustering et de classification sur de multiples espaces ; ceux-ci nous servirons pour exploiter différentes informations concernant les mêmes données et présentes dans des espaces sémantiques différents. Dans un deuxième temps, nous présentons différents algorithmes de réduction de dimensions, indispensables au traitement des quantités de données considérables que nous manipulons tout au long de nos travaux.

Nous établissons ensuite un panorama décrivant les différentes ressources que l'on peut trouver pour les deux tâches d'analyse sémantique que nous étudions. Ces ressources linguistiques sont distinguées en fonction de leur mode de constitution. Nous considérons d'une part les ressources construites manuellement, que ce soit par des linguistes experts ou par des internautes de façon collaborative et à grande échelle, et d'autre part les ressources constituées par l'intermédiaire de calculs automatiques sur de grandes quantités de données.

Les ressources présentées ne couvrent pas l'ensemble de toutes les langues existantes : elles sont généralement beaucoup plus développées pour la langue anglaise. Elles ne peuvent pas non plus être exhaustives sur l'intégralité d'une langue vivante, puisqu'une langue vivante ne cesse d'évoluer. Nous intéressant plus particulièrement au traitement de la langue française, nous étudions les différentes techniques proposées jusqu'ici pour la traduction et l'enrichissement automatiques de telles ressources.

L'étude de ces ressources sémantiques nous mène ensuite aux deux tâches d'analyse sémantique déjà mentionnées. Ainsi nous décrivons les principaux algorithmes proposés dans la littérature concernant d'une part la désambiguïsation lexicale et d'autre part l'annotation en rôles sémantiques.

Pour conclure cet état de l'art, nous passons en revue différents sous-problèmes de Recherche d'Information auxquels nous pensons que l'analyse sémantique peut apporter une amélioration. Les sous-problèmes concernés sont notamment la recherche standard de documents, l'extension et la reformulation de requête, et enfin, les systèmes de réponses à des questions posées en langue naturelle.

## 2.1 Fouille de données

La fouille de données regroupe un ensemble de méthodes d'apprentissage supervisées et non supervisées, pour des tâches telles que la *classification automatique*† [1] ou le *clustering*†. Traditionnellement, ces algorithmes s'appliquent sur des données pour lesquelles on dispose de représentations dans un même espace de caractéristiques (autrement appelées *traits* ou encore *features* en anglais).

La très grande majorité des algorithmes de fouille de données s'appliquent sur de tels espaces, cependant il arrive que certains éléments à classifier ou regrouper disposent de plusieurs représentations différentes, dans des espaces différents. Comment alors exploiter ces différentes données fournissant des informations complémentaires ?

Dans les différentes expériences que nous menons, nous sommes amenée à utiliser des espaces sémantiques multiples. En effet, nous utilisons les espaces sémantiques construits par [Grefenstette 2007] que nous décrirons à la section 3.1.1.2. Dans ce cadre, nous disposons d'un espace distributionnel par type de relation syntaxique ayant servi à calculer les cooccurrences. Nos travaux reposent donc sur l'utilisation de représentations multiples et nous montrerons à la section 3.2.2 que chacune de ces représentations contient des informations différentes et complémentaires. Nous avons donc besoin d'algorithmes spécifiques aux multiples représentations et présentons ici les principales approches existantes de clustering et de classification sur des données multi-représentées.

Une seconde problématique importante dans les travaux que nous menons est liée au passage à l'échelle des calculs sur des données de grande taille. Nous nous intéressons à différentes façons de diminuer le nombre de dimensions des matrices que nous serons amenés à utiliser ainsi qu'à des algorithmes de traitement approximant certaines mesures pour les données de grande taille.

### 2.1.1 Le traitement des données multi-représentées

Afin d'illustrer ce que sont les données multi-représentées, considérons par exemple des éléments *films* qu'il faudrait classer par genre. On peut considérer un espace issu d'une base de données (par exemple Imdb [2]) dans lequel les données factuelles donneraient des traits de nature textuelle (titre, synopsis), symbolique (réalisateurs, acteurs, limitation d'âge) et numérique (année de sortie, durée) et un second espace dans lequel les traits seraient des données purement numériques issues

1. Dans ce manuscrit, tous les mots marqués d'un symbole † sont définis dans le glossaire en annexe A.
2. Internet Movie Database : `http://www.imdb.com`

de l'analyse du flux audio-vidéo lui-même comme dans les travaux de [Rasheed & Shah 2002] : histogrammes de couleur, fréquence de changement de plan et durée moyenne des plans (*average shot length*), perturbation visuelle (*visual disturbance*), fréquence de détection de rires, de pleurs, d'explosions (...) sur la bande-son.

Dans un cas comme celui-là, les deux espaces sont vraiment distincts. Un même élément *film* peut avoir une représentation dans chacun des deux espaces ou n'en avoir qu'une seule dans l'espace issu du flux vidéo. C'est pour ce genre de cas que différents travaux se sont intéressés au clustering sur des représentations diverses appelées selon les auteurs : *agrégation de clusters*, *clustering sur de multiples représentations* ou encore *clustering sur des univers parallèles*.

Comme nous le verrons à la section 2.2.2.4, nous nous intéressons à ce type d'algorithme car ils nous seront utiles pour traiter des mots multi-représentés.

#### 2.1.1.1 Clustering

**Agrégation de clusters** Que ce soit par l'application de différents algorithmes ou par l'utilisation de représentations différentes des mêmes objets, on se retrouve parfois avec différents ensembles de clusters pour un même jeu de données. L'émergence des premiers travaux sur le regroupement ou l'agrégation de ces ensembles (*cluster ensembles* ou *cluster aggregation*) date du début des années 2000. Les travaux les plus reconnus sur le sujet sont ceux de [Strehl & Ghosh 2002], ainsi que ceux de [Fred & Jain 2002], [Topchy *et al.* 2005] et de [Gionis *et al.* 2007].

[Strehl & Ghosh 2002] formalisent de façon très précise le framework des *cluster ensembles* et modélisent le problème comme un problème de maximisation pour lequel ils proposent trois heuristiques de résolution, qu'ils évaluent ensuite sur différents jeux de données. Les trois heuristiques sont fondées sur la constitution d'un hypergraphe où les noeuds représentent les objets et où les clustering initiaux sont représentés par les hyperliens de l'hypergraphe. Le framework permet aussi l'application de cette fusion de clustering en calcul distribué. La distribution des calculs peut se faire soit en répartissant les objets eux-mêmes et leurs représentations, soit en répartissant les représentations des objets, tous les objets étant présents sur tous les lieux de calcul mais avec une seule représentation, ou encore en répartissant les algorithmes de clustering.

Dans le but d'améliorer la qualité des algorithmes de clustering pour des clusters de formes complexes, [Fred & Jain 2002] proposent d'initialiser un grand nombre de petits clusters $k$ fois par l'algorithme *k-means* initialisé de façon aléatoire à chaque fois. Une mesure de similarité est attribuée par vote pour toutes les paires de termes : il s'agit du nombre de fois où deux termes sont attribués au même cluster, divisé par $k$ le nombre de clusterings pré-calculés. Tous les liens inférieurs à un certain seuil sont supprimés et les paires restantes forment un ensemble de graphes connexes constituant les clusters finaux. Les auteurs suggèrent que les paramètres peuvent être déterminés par l'observation des zones stables du nombre de clusters finaux trouvés après avoir fait varier les deux paramètres nombre de clusterings et seuil (au bout d'un temps le nombre de clusters trouvés diverge).

[Gionis *et al.* 2007] étudie le problème des *cluster ensembles* sous le nom de *cluster aggregation* bien qu'il s'agisse exactement du même problème. Tous les cas qui bénéficieraient de ce type de traitement sont clairement décrits. Le problème est pris sous l'angle inverse, il s'agit cette fois de minimiser la somme des distances entre le clustering final et l'ensemble des clustering initiaux. Les auteurs montrent que la minimisation de la distance qu'ils définissent est une restriction du problème de *correlation clustering* défini par [Bansal *et al.* 2004]. Cinq algorithmes minimisant cette distance sont proposés. Ces algorithmes étant fondés sur les algorithmes de résolution du problème de *correlation clustering*, ils sont d'une complexité quadratique. [Gionis *et al.* 2007] proposent donc une étape de pré-traitement et de post-traitement pour pallier ce problème lors du passage à l'échelle sur des grands volumes de données. Tous ces algorithmes sont analysés théoriquement et évalués sur des jeux de données.

Enfin, l'étude complémentaire de [Topchy *et al.* 2004] prouve par deux approches différentes que l'agrégation de clusters par consensus de plusieurs ensembles de clusters converge vers le clustering idéal au fur et à mesure que l'on ajoute de nouveaux ensembles de clusters (à condition que ceux-ci soient meilleurs qu'un regroupement aléatoire).

**Adaptations d'algorithmes de clustering à de multiples représentations** On trouve encore dans la littérature des méthodes plus spécifiques pour l'utilisation d'un même algorithme avec des représentations différentes des données à clusteriser. Parmi celles-ci, [Kailing *et al.* 2004] proposent une adaptation de l'algorithme DBSCAN ([Ester *et al.* 1996]) à de multiples représentations. La prise en compte des différentes représentations est assez simple : il s'agit de considérer le voisinage d'un élément soit comme étant l'intersection des voisinages du même élément dans toutes les représentations, soit comme en étant l'union.

Dans la suite de ces travaux, [Achtert *et al.* 2006] proposent une prise en compte plus fine par l'utilisation d'arbres de combinaisons. Les noeuds terminaux des arbres sont les différentes représentations et les noeuds intermédiaires sont les opérateurs d'union ou d'intersection. Ainsi chaque représentation est fusionnée avec une autre par l'un ou l'autre des deux opérateurs, la fusion étant alors modulable en fonction des représentations. Dans cet article, l'algorithme de base utilisé n'est plus DBSCAN mais OPTICS ([Ankerst *et al.* 1999]) exploitant également la densité des données. Les résultats mettent en valeur l'apport d'une combinaison de deux représentations comparée à l'usage de chacune de ces représentations seule ou dont les caractéristiques seraient concaténées dans un espace commun.

Dans ces articles, l'importance de traiter les représentations d'une manière plus complexe qu'une simple concaténation est argumentée par le fait que les espaces peuvent être dotés de propriétés différentes. On peut par exemple disposer de deux espaces dont un serait *orienté précision* tandis que l'autre serait *orienté rappel*.

Un espace *orienté précision* est défini comme étant un espace dans lequel on aurait :

$$Sim(v1, v2) \Rightarrow Sim(D1, D2)$$

avec D1, D2 deux documents et v1, v2 les signatures de ces documents dans l'espace de représentation. C'est le cas par exemple des modèles de sacs de mots des documents texte.

Un espace *orienté rappel* est défini comme étant un espace dans lequel on aurait :

$$Sim(D1, D2) \Rightarrow Sim(v1, v2)$$

avec D1, D2 deux documents et v1, v2 les signatures de ces documents dans l'espace de représentation. Dans ce cas-là, on peut prendre pour exemple les histogrammes de couleur

Ainsi deux espaces orientés rappel bénéficieraient d'une amélioration de leur précision par une combinaison opérée par l'intersection tandis que deux espaces orientés précision bénéficieraient d'une combinaison par union.

**Univers parallèles ou comment discriminer des clusters globaux localisés dans des représentations différentes** Une autre façon de voir les choses est présentée dans [Patterson & Berthold 2001]. Il s'agit une fois de plus de traiter des données multi-représentées. Dans ce paradigme chaque représentation s'appelle *univers* et les multiples représentations portent l'appellation *univers parallèles*. En revanche, le résultat du clustering dans des univers parallèles ne donne pas des clusters globaux sans distinction de représentation comme précédemment mais des clusters appartenant chacun à un univers en utilisant les informations les plus discriminantes de chaque univers. Ces clusters sont néanmoins disjoints dans l'ensemble global des éléments.

L'approche définit des *neighborgrams* dans chaque univers. Un *neighborgram* est défini comme étant une structure de données résumant le contenu du voisinage d'un élément. Il s'agit d'un histogramme du nombre d'éléments du voisinage en fonction de la distance à l'élément central. Cet histogramme distingue deux classes :la classe positive correspond à la classe de l'élément central tandis que la classe négative correspond aux éléments de toutes les autres classes. Un score est attribué à chaque *neighborgram* de chaque univers en fonction de la concentration des éléments positifs les plus proches de l'élément central. Le meilleur *neighborgram* est validé comme étant un cluster et tous ses éléments positifs sont retirés de chacun des univers. La procédure est alors réitérée jusqu'à épuisement des éléments.

Dans le même courant de pensée on trouve également les travaux de [Wiswedel & Berthold 2007] qui adaptent l'algorithme de *Fuzzy clustering* de [Bezdek 1981] aux univers parallèles. L'algorithme est évalué sur des données artificielles et donne des résultats prometteurs.

Enfin il existe un dernier paradigme proposé par [Bickel & Scheffer 2004] et appelé *Multi-View Clustering*. Il s'agit surtout de co-training (chacune des méthodes se sert de l'autre comme référence pour apprendre). La méthode est valable seulement pour 2 espaces.

#### 2.1.1.2 Classification

Le problème de la classification sur plusieurs représentations est un des cas où les algorithmes de fusion de classifieurs ([Kittler *et al.* 1998], [Kuncheva *et al.* 2001]) sont intéressants. [Duin 2002] apporte une discussion sur la combinaison de classifieurs en répertoriant d'une part les différentes règles fixes de combinaison (produit des scores de confiance, somme, maximum, minimum, médiane) et en décrivant d'autre part un ensemble de traitements pouvant améliorer la combinaison (calibration des sorties des classifieurs de la base, pondération des classifieurs, sélection de sous-ensemble de classifieurs, sélection locale des meilleurs classifieurs, classifieur général de combinaison prenant en entrée les sorties des classifieurs de base). [Duin 2002] conseille aussi l'usage de trois stratégies :

- entraîner les classifieurs de base sur un seul jeu d'apprentissage jusqu'à ce que les estimateurs soient fiables, tout en prenant garde à l'*overtraining*, pour ensuite utiliser une règle fixe de combinaison,
- entraîner faiblement les classifieurs de base sur un seul jeu d'apprentissage et réutiliser ce jeu pour entraîner un classifieur général de combinaison,
- diviser le jeu d'apprentissage en deux parties, entraîner les classifieurs de base sur l'une des parties sans faire spécialement attention à l'overtraining, et entraîner un classifieur général de combinaison sur la deuxième partie.

En revanche, [Duin 2002] déconseille la stratégie communément employée consistant à entraîner les classifieurs de base sans prêter attention à l'overtraining et à entraîner ensuite le classifieur de combinaison sur le même jeu d'apprentissage.

Par ailleurs, [Kriegel *et al.* 2005] proposent une méthode de classification inspirée de l'algorithme k-NN traditionnel ([Cover & Hart 1967]) dans laquelle la combinaison des espaces se fait à l'intérieur même de l'algorithme. La première partie de l'article se concentre sur la réduction du nombre de données d'entraînement (en raison de la grande complexité du k-NN) tandis que la deuxième partie expose l'algorithme de classification avec combinaison intégrée. La méthode utilise des vecteurs de confiance propres à chaque représentation et à chaque terme à classifier pour pondérer la contribution de chaque représentation propre à chaque mot source donné.

Enfin, dans le cadre précis de la désambiguïsation lexicale [Le *et al.* 2007] disposent d'autant de classifieurs qu'ils ont de représentations de leurs éléments. Ils étudient l'impact d'un panel d'opérateurs de fusion fondés sur la *théorie de l'évidence de Dempster-Shafer*† ([Shafer 1976]) et les opérateurs *Ordered Weighted Averaging*† (*OWA*) de [Yager 1988]. Deux opérateurs de type Dempster-Shafer sont définis et utilisés : le premier utilise la règle de *discounting* pour pondérer les *assignations basiques de probabilités (BPA)* issues de chaque représentation ainsi que la *règle de combinaison de Dempster*. Le second opérateur utilise également le *discounting* mais l'opérateur de combinaison utilisé est la moyenne des BPA. Les opérateurs OWA étudiés sont les opérateurs minimum des probabilités conditionnelles, maximum, médiane, vote majoritaire, vote majoritaire flou. L'opérateur de fusion utilisant la règle de combinaison de Dempster donne en moyenne de meilleurs résultats que tous les autres opérateurs proposés et surpassent également d'autres opérateurs de combinaison proposées par la littérature (y compris le vote majoritaire couramment utilisé en fusion de classifieurs de désambiguïsation lexicale).

#### 2.1.1.3 Mesures de similarité

Plus généralement, certains travaux s'intéressent aussi à l'estimation de similarité globale d'objets multi-représentés. C'est le cas de [Kriegel *et al.* 2008a] et [Kriegel *et al.* 2008b] qui soulignent le fait que certaines représentations peuvent ne pas désambiguïser la similarité ou dissimilarité de deux objets tandis que d'autres représentations seront plus significatives sur ces objets donnés. L'idée générale de ces approches consiste à pondérer le pouvoir discriminatoire de chacune des représentations pour deux objets donnés : deux objets ayant la même distance dans une représentation donnée partagent-ils réellement le même degré de similarité ou de dissimilarité globale ? Ces études proposent un estimateur joint de similarité et non-dissimilarité et un algorithme permettant aux estimateurs d'apprendre leurs paramètres en fonction des données de représentation. Une expérience de recherche par similarité montre un gain de pertinence comparé à des approches plus standard que sont la moyenne des représentations ou les résultats issus de la meilleure représentation.

#### 2.1.1.4 Bilan

Différentes tendances et appellations sont apparues sur la problématique commune qu'est la fouille de données sur des objets possédant plusieurs représentations. Les méthodes diffèrent les unes des autres sans qu'aucune étude comparative soit menée sur le sujet. La seule conclusion que nous pouvons tirer de ces différentes études est qu'il s'avère important de prendre en compte de façon indépendante les informations de similarité ou de dissimilarité contenues dans chacune des représentations. En effet, chacune des représentations possède un pouvoir discriminatoire variant en fonction des éléments de l'univers. Certaines représentations sont plus aptes à discriminer des éléments donnés, tandis que d'autres représentations discriminent plus aisément d'autres éléments. Les algorithmes appliqués à de telles représentations doivent être capables d'ordonner le pouvoir

discriminatoire de chaque représentation sur chaque sous-ensemble d'éléments, de façon à ne prendre en compte que l'information la plus fiable de chaque représentation.

### 2.1.2 La réduction de dimensions

Les algorithmes classiques de fouilles de données sont contraints par des complexités dépendant du nombre de dimensions des vecteurs traités. Cette complexité (au minimum quadratique pour la plupart des algorithmes standards) est encore aujourd'hui un phénomène bloquant pour la manipulation de vecteurs de très grandes dimensions en un temps raisonnable. À cela, plusieurs solutions peuvent répondre. On peut trancher abruptement en déterminant arbitrairement un facteur qui rendrait une caractéristique non pertinente (fréquence d'occurrence ou autre). Une autre solution consiste à utiliser certains algorithmes aléatoires qui ne font qu'approximer les résultats souhaités. Enfin, on peut aussi exploiter la faible densité des matrices ou la redondance d'information corrélée y figurant pour réduire leur taille tout en conservant le maximum d'information intacte. C'est cette dernière alternative que nous appelons réduction de dimensions et dont nous présentons ici les principales méthodes existantes.

Nous évoquerons dans un premier temps les techniques d'Analyse en Composantes Principales et de Décomposition en Valeurs Singulières. Nous étudierons ensuite le détail de la définition d'une famille de hachage permettant d'obtenir des signatures conservant la similarité cosinus des vecteurs initiaux ainsi que la démonstration de cette propriété. Enfin, nous présenterons un algorithme de recherche rapide de plus proches voisins utilisant les signatures obtenues avec le hachage précédemment décrit.

#### 2.1.2.1 Analyse en Composantes Principales et Décomposition en Valeurs Singulières

L'Analyse en Composantes Principales (*ACP* ou *PCA* en anglais) est une méthode consistant à transformer des variables corrélées en nouvelles variables indépendantes les unes des autres. Ces nouvelles variables non corrélées sont appelées *composantes principales*. A l'issue d'une analyse en composantes principales, ces composantes (autrement appelées axes) sont ordonnées de la plus informative à la moins informative. On peut alors pour compresser les données d'une matrice, ne conserver que les $n$ premières pour conserver le maximum d'information.

La décomposition en valeurs singulières est également une méthode de factorisation de matrices. Dans une décomposition en valeurs singulières $M = U.\Sigma.V$, les deux matrices U et V sont des matrices contenant des bases orthonormées et $\Sigma$ est une matrice diagonale contenant les valeurs singulières. A l'issue de cette décomposition, les vecteurs de U représentent les directions de plus grande variation de l'ensemble et les valeurs diagonales de $\Sigma$ représentent la pondération de ces directions. Dans l'objectif de réduire les dimensions d'une matrice, on peut alors considérer uniquement les $n$ premières valeurs diagonales de $\Sigma$ et assigner les autres à 0 pour ainsi reconstituer une matrice $M'$ qui ne contient que l'information la plus dominante de M.

Pour les détails mathématiques de ces deux méthodes de compression de matrices, nous invitons le lecteur à lire le rapport de [Madsen *et al.* 2004].

#### 2.1.2.2 Hachage Sensible à la Localité (LSH)

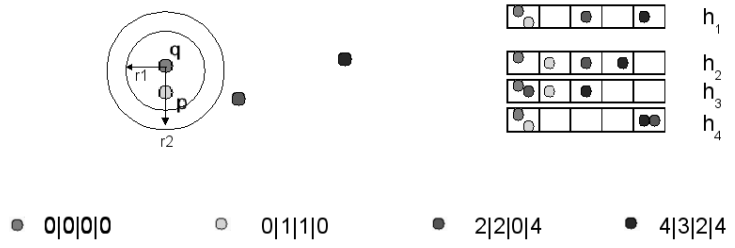



FIGURE 2.1 – Principe du Hachage sensible à la localité

Le but d'une fonction de hachage est d'obtenir une empreinte plus petite que l'élément initial. Pour deux vecteurs similaires $\overrightarrow{u}$ et $\overrightarrow{v}$ de l'espace initial, on souhaite pouvoir retrouver cette similarité entre les empreintes hachées $h(\overrightarrow{u})$ et $h(\overrightarrow{v})$. La Figure 2.1 illustre ce principe. Les mots représentés par les points p et q sont proches dans l'espace initial. Supposons que nous connaissions une famille de fonctions de hachage $h_1$, $h_2$, $h_3$, $h_4$, attribuant chacune une valeur comprise entre 0 et 4 à chacun des vecteurs d'origine. Ces valeurs sont représentées par les différentes cases de la figure, la première case correspondant à la valeur 0. Les signatures résultantes du hachage de p et q sont : *signature(q) = 0/0/0/0* et *signature(p) = 0/1/1/0*. Dans le cas où la probabilité de collision (fait que deux vecteurs se voient attribuer la même valeur par une des fonctions de hachage) est forte pour deux vecteurs d'origine proches, on obtient effectivement des signatures hachées proches comparées aux signatures hachées des éléments lointains (2|2|0|4 et 4|3|2|4).

À partir des travaux de [Indyk & Motwani 1998], [Charikar 2002] définit une famille de fonctions LSH produisant des empreintes sur lesquelles on peut calculer une approximation de la similarité cosinus beaucoup plus rapidement que dans l'espace d'origine. De plus, [Ravichandran *et al.* 2005] montrent que ce hachage est particulièrement adapté pour mettre en place une méthode de recherche rapide de plus proches voisins approximatifs. Nous reprenons ici les grandes lignes de la méthode de hachage.

On tire $d$ vecteurs unitaires $\overrightarrow{r}$ selon une distribution gaussienne. Ce tirage assure une répartition équidistribuée sur l'hypersphère unitaire. Soit une famille de fonctions définies par :

$$h_{\overrightarrow{r}}(\overrightarrow{u}) = \begin{cases} 0 \; if \; \overrightarrow{r}.\overrightarrow{u} \geq 0 \\ 1 \; if \; \overrightarrow{r}.\overrightarrow{u} < 0 \end{cases} \tag{2.1}$$

La Figure 2.2 illustre ce hachage sur une vue simplifiée à deux dimensions. Les vecteurs $\overrightarrow{r_1}$, $\overrightarrow{r_2}$ et $\overrightarrow{r_3}$ définissent des hyperplans équirépartis sur l'hypersphère. Les vecteurs $\overrightarrow{u}$, $\overrightarrow{v}$ et $\overrightarrow{w}$ sont les vecteurs de notre espace initial. En suivant la formule de hachage donnée par 2.1, on obtient les signatures présentées sur la figure. On remarque que pour un petit nombre $d$ de clés de hachage (ce nombre est réduit uniquement dans un but illustratif) , les vecteurs $\overrightarrow{v}$ et $\overrightarrow{w}$ qui étaient très proches dans l'espace d'origine, ont deux signatures identiques. Ici, la valeur obtenue par chacune des clés de hachage est identique, la collision (fait d'obtenir une valeur identique) est systématique. Un nombre plus grand de clés permettraient d'obtenir des hyperplans séparant $\overrightarrow{v}$ et $\overrightarrow{w}$ et donc de garder une collision forte mais non systématique. Les signatures obtenues seraient alors similaires.

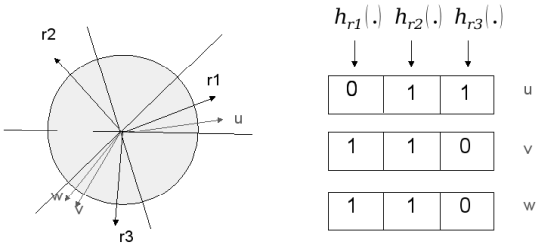



FIGURE 2.2 – Vue en 2D de la méthode de LSH

**Démonstration** Soient deux vecteurs $\overrightarrow{u}$ et $\overrightarrow{v}$. La probabilité de tirer un vecteur aléatoire $\overrightarrow{r}$ définissant un hyperplan qui les séparera est égale à :

$$Pr[h_{\overrightarrow{r}}(\overrightarrow{u}) \neq h_{\overrightarrow{r}}(\overrightarrow{v})] = \theta(\overrightarrow{u}, \overrightarrow{v})/\Pi \tag{2.2}$$

Sur un nombre $d$ de vecteurs tirés aléatoirement on peut mesurer cette probabilité. En effet, la probabilité qu'un hyperplan tiré aléatoirement ait séparé les deux vecteurs originaux $u$ et $v$ est la probabilité que cet hyperplan ait donné un bit différent pour les deux résultats du hachage de $u$ et $v$. La formule 2.3 nous donne cette probabilité :

$$Pr[h_{\overrightarrow{r}}(\overrightarrow{u}) \neq h_{\overrightarrow{r}}(\overrightarrow{v})] = distance\_de\_Hamming(\overrightarrow{u}, \overrightarrow{v})/d \tag{2.3}$$

En combinant 2.2 et 2.3 on obtient donc l'approximation :

$$\cos(\theta(\overrightarrow{u}, \overrightarrow{v})) \approx \cos(distance\_de\_Hamming(\overrightarrow{u}, \overrightarrow{v})/d * \Pi) \tag{2.4}$$

$distance\_de\_Hamming(\overrightarrow{u}, \overrightarrow{v})/d * \Pi$ variant entre 0 et $\Pi$, la distance de Hamming est elle-même une distance conservant (à l'inverse puisqu'il s'agit d'une distance) les similarités de la similarité Cosinus dans l'espace original.

#### 2.1.2.3 Recherche des plus proches voisins approximatifs (A-NN)

Une recherche rapide du plus proche voisin approximatif dans un espace muni d'une distance de Hamming a été proposée par [Charikar 2002] et reprise par [Ravichandran *et al.* 2005]. Elle permet d'exploiter les vecteurs issus du hachage LSH.

La méthode consiste à tirer aléatoirement $p$ permutations de $d$ éléments ($d$ étant la taille de la signature LSH). Pour chaque permutation, on permute les signatures bit à bit, on procède à un tri lexicographique de tous les éléments et on garde les $B$ plus proches éléments des $n$ éléments source dont l'approximation du cosinus est inférieur à un certain seuil. Cela fonctionne car une signature permutée est une représentation valide du vecteur d'origine. C'est en soi une signature par une famille de hachage. Il suffit ensuite de procéder au tri des $B.P$ éléments obtenus lors des étapes intermédiaires pour obtenir $k < B.P$ plus proches voisins pour une complexité beaucoup moins grande.

#### 2.1.2.4 Bilan

La technique de hachage décrite ainsi que la méthode de recherche des plus proches voisins associée sont particulièrement adaptées à nos données. En effet, la mesure cosinus est une très bonne métrique de similarité sémantique ([Curran 2004]) et la diminution de la complexité de calcul de cette mesure est pour nous essentielle.

Ces algorithmes de réduction de dimensions et de calcul approximatif se montreront indispensables dans la manipulation des espaces vectoriels que nous introduirons à la section 2.2.2.3 dans lesquels les éléments sont des mots et la distance entre ces mots représente la notion de dissimilarité de sens.

### 2.1.3 Conclusions

Les différents algorithmes répertoriés pour le traitement des données multi-représentées nous seront très utiles pour exploiter pleinement nos représentations des mots. En effet, chacune des ressources que nous allons voir dans la section suivante propose une caractérisation différente des mêmes mots de la langue. Nous pensons qu'un usage unifié des différents types de ressources bénéficierait grandement d'algorithmes de fusion ne conservant que les résultats pour lesquels la confiance peut être estimée élevée et négligeant ceux pour lesquels la confiance est faible. Nous tenterons de valider l'intérêt de cette combinaison dans les chapitres à venir.

Les algorithmes de réduction de dimensions sont également indispensables à la réalisation de nos travaux. Nous verrons dans la section suivante que la taille des données traitées nécessite une réduction de dimensions importante et que le hachage *LSH* y est particulièrement adapté.

## 2.2 Les ressources linguistiques à l'usage de la sémantique

La communauté du Traitement Automatique des Langues (TAL) dispose d'un certain nombre de ressources linguistiques. Certaines d'entre elles sont d'anciennes ressources qui furent numérisées, mais on constate également l'émergence de nouvelles ressources produites par des linguistes, des experts métiers et des informaticiens, suite aux besoins ressentis par la communauté du TAL à

l'avènement de l'ère numérique. Nous restreignons ici le cadre de notre étude aux ressources utiles aux deux tâches d'analyse sémantique identifiées précédemment (Section 1.4) que sont la *désambiguïsation lexicale* et l'*analyse en rôles sémantiques*.

Nous distinguons ici les ressources manuelles, constituées par des humains, allant d'un comité d'experts à l'ensemble des internautes dans le cas de ressources collaboratives, des ressources automatiques qui nécessitent des calculs effectués par une machine.

### 2.2.1 Ressources manuelles

Nous regroupons sous l'appellation ressources manuelles toute ressource constituée par des humains sans calcul machine. Cela fait donc référence à la fois aux ressources construites sur des fondements théoriques mis en œuvre après de longs mois de discussions par le consensus commun d'un comité d'experts ainsi qu'aux ressources collaboratives pour lesquelles tout un chacun peut contribuer à l'enrichissement quotidien de ladite ressource et corriger les errements de ses prédécesseurs lorsque cela est jugé nécessaire. D'aucuns se glaceront d'horreur à la vue d'un tel regroupement. Nous considérons pour notre part que les qualités intrinsèques aux premières (telles que l'exactitude théorique des informations qui y sont répertoriées) sont l'idéal vers lequel tendent les secondes au fur et à mesure du temps.

#### 2.2.1.1 Ressources syntaxiques

#### 2.2.1.2 Ressources lexicographiques et ontologies

La définition d'une ontologie n'est pas constante d'un lexique à un autre. On peut cependant regrouper quelques points communs. Une ontologie est un réseau lexical. On peut la modéliser par un graphe dans lequel les mots ou concepts sont les sommets du graphe, tandis que les arêtes représentent des relations particulières entre ces concepts. Chaque concept est doté d'une définition. Parmi les relations présentes dans une ontologie, on trouve presque systématiquement les relations *taxinomiques* (ou *taxonomiques* encore appelées relations de *subsomption*). Ces relations exprime le fait qu'un mot appartient à la catégorie défini par le second mot, comme dans l'exemple ***chat** est un **animal***. On dit alors que **chat** est un *hyponyme* de **animal** et qu'**animal** est un *hyperonyme* de **chat**. Dans une ontologie, il existe généralement d'autres types de relations (sinon il s'agirait plus d'une *taxonomie* que d'une ontologie), de type sémantique ou métier

Il existe deux types d'ontologies : traditionnellement les ontologies ne se veulent pas exhaustives sur l'ensemble du langage, mais les plus complètes possibles sur un domaine métier. Par extension, on appelle aussi ontologie un réseau lexical possédant les propriétés décrites ci-dessus et ayant pour ambition de couvrir l'ensemble du langage. Cependant, il est recommandé de le préciser par un qualificatif supplémentaire comme pour les ressources suivantes :

- *WordNet* pourtant défini comme une base de donnée lexicale ([Miller 1995]) est parfois qualifié d'ontologie **lexicale**, **lexicologique** ou encore **linguistique générale**
- La *Top Ontology* ainsi que les *Base Concepts* d'EuroWordNet ([Vossen 1998]) sont considérés comme des **ontologies générales**.
- *SUMO* est l'acronyme de *Suggested* ***Upper*** *Merged Ontology* [Niles & Pease 2001]. SUMO est une ontologie généraliste concernant les concepts les plus hauts de l'échelle taxonomique.

**WordNet** Développé dès 1995 par l'Université de Princeton ([Miller 1995], [Fellbaum 1998]), WordNet est aujourd'hui la ressource libre la plus exploitée par les chercheurs intéressés par la sémantique et ses thématiques connexes. WordNet est un réseau lexical couvrant l'ensemble du vocabulaire anglophone et un bon nombre d'entités nommées de type lieu, personne, organisation.

La structure de WordNet est un peu différente de celle d'un dictionnaire ou d'une taxonomie classique. En effet, les synonymes sont regroupés entre eux pour former ce qu'on appelle des *synsets*, entités unitaires de WordNet. Ainsi, un même mot peut appartenir à différents synsets, qui constituent différents sens ou usages du même mot. Ce sont les synsets eux-mêmes qui comportent une définition textuelle appelée *gloss* et qui sont reliés entre eux par des relations sémantiques. Les principales relations que l'on trouve pour les noms sont les relations suivantes :

**antonymie** on parle d'antonymie lorsqu'un concept décrit le contraire d'un autre. *petit* est l'antonyme de *grand* et vice-versa ;

**hyponymie/hyperonymie (subsomption)** Il s'agit de la relation présente dans les taxonomies déjà illustrée plus haut ;

**méronymie/holonymie** On parle de méronymie lorsqu'un concept fait partie d'un autre, e.g. un *guidon* est un méronyme de *vélo* et *vélo* est un holonyme de *guidon*.

Pour les verbes on trouve également des relations de troponymie et d'implication (*entailment*). Une étude plus détaillée des relations trouvées dans WordNet est fournie à la section 3.1.3.1.

La dernière version de WordNet (WordNet 3.0) contient plus de 110 000 expressions nominales (dont environ la moitié ne sont formées que d'un seul mot), plus de 20 000 verbes, plus de 10 000 adjectifs et près de 5000 adverbes.

**EuroWordNet** L'objectif du projet EuroWordNet ([Vossen 1998]) fut de constituer un équivalent du WordNet original de Princeton pour plusieurs des langues européennes. Une structure de réseau lexical est construite pour chaque langue pour tenir compte des spécificités linguistiques de chacune. Il existe cependant une structure commune partagée par tous les réseaux lexicaux. Cette structure représente l'ensemble des concepts et relations possibles exprimant ainsi tous les concepts présents dans au moins une langue mais pas nécessairement dans toutes. Chaque synset de chaque langue est relié par une relation d'équivalence à un noeud de cette structure afin de pouvoir mettre en correspondance les différents synsets des différentes langues. La structure commune comprend également

une ontologie de haut niveau (*Top Level Ontology*) ainsi qu'une hiérarchie de domaines.

Tous les WordNets sont constitués à partir de ressources existantes. Pour la plupart des langues, une approche par fusion est menée : les entrées et différentes relations sont constituées à partir de ressources externes pour être ensuite reliées par des liens d'équivalence à WordNet 1.5. Pour l'espagnol, l'approche adoptée est extensive : les mots sont choisis à partir de WordNet 1.5, traduits par dictionnaires bilingues et ensuite filtrés pour résoudre les problèmes de polysémie. Le choix de l'approche est fortement lié à la qualité des ressources existant pour chacune des langues. Chaque fragment de ressource constitué par ce biais est vérifié par un panel d'utilisateurs. Ce projet a produit une ressource française contenant un peu plus de 20 000 synsets pour près de 140 000 dans la version 1.5 du WordNet anglais utilisé. Malheureusement EuroWordNet est distribué sous une licence propriétaire restrictive[3].

**SUMO** Sumo a été construit avec pour ultime but de regrouper toutes les ontologies existantes sous l'égide d'une seule et unique *Upper Ontology*. Le projet est initié en 2001 par une collaboration interdisciplinaire entre l'ingénierie, la philosophie et les sciences de l'information.

SUMO est effectivement aujourd'hui une des deux plus grandes ontologies connues. Elle regroupe entre autres les ontologies de haut niveau proposées par [Sowa 2000] et [Russell & Norvig 2003], des extraits de la DBPedia ([Auer *et al.* 2008]), un mapping intégral avec WordNet, ainsi qu'un très grand nombre d'autres ontologies spécialisées.

Ceci constitue un total de 20 000 termes et 70 000 axiomes (équivalent des relations). Ces chiffres ne tiennent pas compte d'une grande partie des entrées de WordNet. Par exemple, à la classe terminale *amphibian* de SUMO, sont rattachées 164 entrées de WordNet (littéraux) elles-mêmes hiérarchisées entre elles.

SUMO appartient à l'IEEE et est distribuée sous licence GNU/GPL.

**CYC** ([Lenat 1995]) décrit CYC, une base de connaissances de sens commun dont la constitution manuelle a commencé en 1984. Depuis cette date, plus de 500 000 concepts décrivant le monde tels que les humains le perçoivent ont été entrés, ainsi que plus de 5 000 000 d'assertions décrivant ces concepts et 26 000 relations les reliant. Cyc répertorie aussi des contextes dans lesquels une assertion peut être vraie ou ne pas l'être. Par exemple, dans le contexte d'*obscurité totale*, l'assertion *Vous ne voyez rien* est vrai, ce qui n'est pas le cas dans un contexte d'*éclairage*. Les modèles de raisonnement de CYC ne sont ni de la logique binaire, ni de la logique floue arithmétique au sens traditionnel du terme, mais la logique est assurée par des méta-assertions de type "*Assertion A* est moins probable que *Assertion B*". CYC se veut exhaustif sur toute la connaissance qui n'est ordinairement écrite nulle part mais que tout un chacun apprend dans ses plus jeunes années, comme par exemple :

- *On ne peut pas se rappeler d'événements qui ne se sont pas encore produits.*
- *On peut habituellement voir le nez d'une personne mais pas son coeur.*

3. `http://www.elda.org/catalogue/fr/text/M0015.html`

Deux versions de CYC ont été rendues publiques. OpenCyc est une version comprenant majoritairement des assertions de nature taxonomique, tandis que ResearchCyc, mis à disposition de la communauté scientifique à des fins de recherche, contient un certain nombre d'assertions plus complexes. OpenCyc est distribué sous licence Apache 2, permettant la copie, la distribution et l'usage dans des buts commerciaux ou non. ResearchCyc est distribué sous demande lorsque l'égibilité peut être vérifiée.

Les concepts de CYC ont été reliés à ceux d'autres ressources comme WordNet ([Reed & Lenat 2002]), la *Top-Ontology* d'EuroWordNet ([Kiryakov & Simov 2000]), DBPedia ([Alilaghatta 2006]), Geonames ou encore Yago [4]. Aujourd'hui, certains travaux comme ceux de [Sarjant *et al.* 2009] visent à agrandir encore la base de données de CYC avec d'autres données, comme dans ce cas-là, celles contenues dans la ressource collaborative Wikipédia (cf. ci-dessous).

**Wiktionnaire** Le Wiktionnaire [5] est un ensemble de dictionnaires multilingues collaboratifs, hébergés et créés par la fondation Wikimedia, sous licence GFDL (GNU Free Documentation license). Il est écrit au format MediaWiki, un format semi-structuré dont les standards laissent les contributeurs très libres sur la façon d'y écrire les données. Il existe un dictionnaire multilingue par langue d'édition et la syntaxe de structure change en fonction de celle-ci.

Ces dictionnaires répertorient différentes propriétés linguistiques des mots. On trouvera entre autres les informations sémantiques suivantes : les définitions des mots selon leurs différents sens, leurs synonymes et certaines autres relations sémantiques, ainsi que leurs traductions dans d'autres langues. Un mot de la langue d'édition du dictionnaire peut constituer une entrée et être traduit dans sa section traduction. Mais un mot d'une autre langue que la langue d'édition peut aussi constituer une entrée et dans une section traduction être traduit vers la langue d'édition. Un point intéressant à noter est la distinction fréquente des traductions selon les différents sens du mot source.

**Wikipédia** Nous référençons également la ressource Wikipédia [6], ressource collaborative antérieure au Wiktionnaire, également fondée et hébergée par la fondation Wikimedia sous licence GFDL. Cette ressource bien connue du commun des jeunes (et moins jeunes) mortels a pour vocation de prolonger les projets antérieurs de dictionnaires universels de [Chambers 1728] ou de [Diderot & D'Alembert 1751] (autrement appelés respectivement *Cyclopædia* [7] ou *Dictionnaire universel des arts et des sciences* et *Encyclopédie* [8] ou *Dictionnaire raisonné des sciences, des arts et des métiers*), pour ne citer que les travaux modernes (!) où les articles ont commencé à être classés selon différentes branches thématiques. Ainsi la Wikipédia se donne pour objectif de décrire et répertorier le plus possible de connaissance humaine grâce à la participation de tous les internautes le souhaitant.

4. Pour une liste complète, voir : `http://wiki.dbpedia.org/Downloads32#h69-1`
5. `http://wiktionary.org`
6. `http://wikipedia.org`
7. Version numérisée disponible sur `http://digicoll.library.wisc.edu/HistSciTech/subcollections/CyclopaediaAbout.html`
8. Version numérique disponible sur `http://diderot.alembert.free.fr/`

En ce qui nous concerne, cette ressource s'avère intéressante à des fins de désambiguïsation car elle répertorie différents articles lorsque qu'un mot fait référence à différentes entités. Par exemple, la page concernant le mot référent *orange*[9] contient des descriptions et des liens vers une trentaine de références différentes allant du fruit à l'*agent orange* en passant par la couleur éponyme, l'entreprise éponyme, une quinzaine de références géographiques, quelques patronymes et encore quelques autres entités tels que film, bateau ou carte de transport portant ce nom.

#### 2.2.1.3 Ressources pour l'annotation en rôles sémantiques

Concernant l'annotation en rôles sémantiques ou rôles thématiques, plusieurs ressources sémantiques ont été construites, s'essayant à décrire de façon plus ou moins exhaustive les rôles que peuvent jouer les différents éléments d'un texte pour chaque mot considéré comme l'expression d'une action (verbe, nom, adjectif...). Nous répertorions ici les ressources les plus connues. Nous ne faisons référence qu'à des ressources purement sémantiques, excluant ainsi des ressources plus syntaxiques comme Dicovalence un dictionnaire de valence des verbes français ([Eynde (van den) & Mertens 2003] et [Mertens 2010]) ou les tables du Ladl autrement appelé lexique-grammaire ([Leclère 2005]), bien que l'on puisse également inférer une information sémantique à partir de ces dernières.

VerbNet [Schuler 2005] décrit un ensemble de rôles communs à tous les verbes (*agent, theme, location, ...*) et attribue à chacun des verbes de son dictionnaire un sous-ensemble de ces rôles. VerbNet contient actuellement plus de 3 700 verbes.

Dans PropBank [Kingsbury & Palmer 2002], on trouve un ensemble de rôles communs à tous les verbes (Arg-0, Arg-1, Arg-2, Arg-LOC, etc.). L'argument Arg-1 correspond systématiquement au rôle d'*agent* du verbe quand il y en a un. L'argument Arg-1 correspond très souvent au rôle de *patient* quand il n'est pas à la fois *agent*. Cependant ces rôles sont plus spécifiquement définis pour chaque verbe. Par exemple, pour le verbe anglais *to declare*, l'argument Arg-0 est défini comme le *declarer* et l'argument Arg-1 comme un *item being described + description*.

La théorie des frames sémantiques de [Fillmore 1976] est fondatrice de la ressource FrameNet ([Baker *et al.* 1998]). Développée dans ce paradigme, elle définit un certain nombre de cadres appelés *frames* décrivant chacun une situation précise susceptible d'apparaître dans notre perception du monde. Chaque frame est ainsi définie par rapport à un ensemble de rôles qui lui est spécifique. Ces rôles sont appelés *frame elements - FE*. FrameNet répertorie aussi des prédicats (verbes, noms, adjectifs et même adverbes) appelés unités lexicales (*Lexical Units - LU*) déclenchant ces situations, autrement dit les prédicats régissant ces ensembles de rôles. Par exemple la frame */Ingestion/* comprend des rôles tels que *Ingestor*, *Ingestibles*, *Degree*, *Duration*, *Instrument*, et des unités lexicales *breakfast.v*, *consume.v*, *dine.v*, *drink.v*, *sip.v*, *sip.n*, etc. La base contient actuellement plus de 10 000 unités lexicales.

9. http://fr.wikipedia.org/wiki/Orange

L'annotation en rôles sémantiques (*Semantic Role Labeling*) consiste à faire correspondre les différents syntagmes des phrases d'un texte avec les rôles sémantiques correspondants décrits dans ces ressources. Le tableau 2.1 présente les caractéristiques des trois ressources principales anglaises, ainsi qu'un exemple d'annotation associé. Parmi ces ressources, les rôles sémantiques peuvent être peu nombreux et communs à toutes les situations (5 arguments principaux dans PropBank) ou nombreux et spécifiques à chaque situation (environ 250 rôles dans FrameNet).

| Ressource | Nombre de rôles | Exemple de situation | Exemple de phrase |
|---|---|---|---|
| VerbNet | 21 | keep-15.2 : agent, theme, location | He *[agent]* **left [keep-15.2]** the car *[theme]* in the park *[location]*. |
| PropBank | 5 | leave.02 : Arg0, Arg1, Arg2, Argm-TMP | He *[Arg0]* **left [leave.02]** the car *[Arg1]* in the park *[Arg2]*. |
| FrameNet | 250 | Departing : Source, Theme, Place, Circumstances, etc... | He *[Theme]* **left [Departing]** the car *[Source]* in the park *[Place]*. |

TABLE 2.1 – Comparaison des différents rôles sémantiques

**FrameNet** FrameNet nous a paru la ressource la plus intéressante en raison de deux propriétés dont elle est la seule à disposer. En effet, d'une part FrameNet regroupe différents prédicats sous une même frame, ce que n'est pas le cas des autres ressources décrites, d'autres part, ces frames (ainsi que les rôles associés) sont reliées entre elles par des relations sémantiques (héritage, subsomption, cause, précédence, nécessité, exclusion,...). Ces deux caractéristiques permettent ainsi une certaine généralisation et peuvent être utiles à la fois pour les systèmes de Q/R et pour les tâches de raisonnement. Nous verrons plus en détails à la section 6.3.3, comment ces deux propriétés peuvent être exploitées pour améliorer les systèmes de Q/R.

Nous intéressant plus particulièrement à FrameNet, nous décrivons en détails son origine et son contenu. Le projet FrameNet débuta à Berkeley en 1997 ([Baker *et al.* 1998]). La ressource FrameNet telle qu'on la connaît aujourd'hui a été construite pour la langue anglaise à partir du British National Corpus (BNC). Une équipe de lexicographes a annoté 150 000 phrases du BNC en se fondant sur la théorie linguistique des cadres développée par [Fillmore 1968], [Fillmore 1976], afin de répertorier un certain nombre de *cadres* (*frames*), les *rôles* (*frame elements*) qui leur sont associés, des exemples réalisant ces rôles, ainsi qu'un certain nombre de propriétés de valence syntaxique (cadre de sous-catégorisation).

Les étapes de constitution de la base de données FrameNet furent les suivantes :

**Préparation** génération manuelle de descriptions initiales des patterns sémantiques et morphosyntaxiques qui seront utilisés dans les requêtes d'extraction et donc dans les annotations du corpus ;

**Extraction du sous-corpus** extraction automatique (ou manuelle/interactive en cas d'échec) d'exemples de phrases adaptés ;

**Annotation** annotation manuelle des constituants pertinents trouvés dans le corpus ;

**Écriture de l'entrée FrameNet** construction manuelle d'une base de données de représentations sémantiques lexicales issues des annotations, contenant à la fois les lemmes correspondant aux prédicats appelés *Lexical Units* (LUs) et les syntagmes correspondant aux rôles appelés *Frames Elements*.

Au final le projet a généré trois ressources linguistiques très fortement reliées entre elles.

La première ressource est un lexique correspondant à l'index des frames par les prédicats. Il contient des données textuelles à usage d'un utilisateur humain, des formules capturant les formes morphosyntaxiques dans lesquelles les frames peuvent apparaître autour des mots, des liens vers des exemples annotés du corpus, ainsi que des liens vers la frame correspondante dans la base de donnée FrameNet, ou vers d'autres ressources linguistiques comme WordNet ou COMLEX [Grishman *et al.* 1994].

La deuxième ressource est la base de données FrameNet elle-même contenant toutes les frames. Chaque frame est définie par une définition décrivant la situation et les différents rôles intervenant dans cette situation. Nous prendrons pour exemple la frame *Ingestion* dont la définition est la suivante : *An Ingestor consumes food or drink (Ingestibles), which entails putting the Ingestibles in the mouth for delivery to the digestive system. This may include the use of an Instrument. Sentences that describe the provision of food to others are NOT included in this frame* [10]

Chaque frame contient une énumération de ses prédicats (unités lexicales) et leur définition. Pour la frame *Ingestion*, nous retrouvons quelques exemples de prédicats : *breakfast.v*, *consume.v*, *dine.v*, *drink.v*, *sip.v*, *sip.n*. Les frames contiennent également les rôles associés et leur définition. Par exemples les rôles principaux de la frame *Ingestion* sont définis de la façon suivante :

**Ingestibles** The *Ingestibles* are the entities that are being consumed by the *Ingestor*. [11]

**Ingestor** The *Ingestor* is the person eating or drinking. [12]

Enfin, les frames contiennent également les fréquences d'usage syntaxiques issues du BNC pour chaque unité lexicale. Celles-ci sont données à titre d'exemple pour l'unité lexicale *eat.v* dans le tableau 2.2. La première ligne de chaque occurrence spécifie le type de syntagme réalisant le rôle, tandis que la deuxième ligne spécifie ce qui est défini dans FrameNet comme étant la fonction grammaticale du syntagme (*Grammatical Function - GF*).

10. Un *Ingestor* consomme de la nourriture ou une boisson (*Ingestibles*), ce qui nécessite de mettre les *Ingestibles* dans la bouche pour les amener au système digestif. Ceci peut induire l'utilisation d'un *Instrument*. Les phrases qui décrivent l'apport de nourriture à d'autres (personnes ou animaux) ne concernent pas cette frame.
11. Les *Ingestibles* sont les entités consommées par l'*Ingestor*.
12. L'*Ingestor* est la personne qui mange ou qui boit.

| Nombre d'annotations | Patrons | |
|---|---|---|
| Total 26 | Ingestibles | Ingestor |
| (1) | CNI[1] | NP[2] |
| | – | Ext[3] |
| (4) | INI[4] | NP |
| | – | Ext |
| (1) | NP | PP[by][5] |
| | Ext | Dep[6] |
| (3) | CNI | NP |
| | – | Ext |
| (1) | NP | CNI |
| | Obj[7] | – |
| (16) | NP | NP |
| | Obj | Ext |

TABLE 2.2 – Valence syntaxique des réalisations de rôles pour l'unité lexicale *eat.v*

[1] *CNI : Constructional Null Instanciation* - Le constituant est omis en raison de la construction grammaticale de la phrase (*e.g.* ***Cook*** *on low heat until done. [CNI Food]*)

[2] *NP : standard Noun Phrase* - Syntagme nominal

[3] *Ext : External Argument* - Qualifie un syntagme externe au syntagme maximal couvrant le prédicat. Ce syntagme peut être employé en tant que sujet d'un verbe prédicat (*e.g. [The physician]* ***performed*** *the surgery.*), ou bien en tant que sujet, objet direct ou indirect d'un autre verbe mais contrôlant le sujet du verbe prédicat (*e.g. They persuaded [the doctor] to* ***treat*** *me.*), ou enfin en tant que dépendant d'un nom et contrôlant le verbe prédicat (*e.g. Today's decision [by the Court] to* ***approve*** *our request (...)*)

[4] *INI : Indefinite Null Instanciation* - Le constituant est omis en raison de l'usage intransitif d'un verbe ordinairement transitif (*e.g. He* ***takes*** *and never give back. [INI Theme]*)

[5] PP[by] : Prepositional Phrase - syntagme prépositionnel utilisant la préposition *by*

[6] *Dep : Dependent* Fonction grammaticale attribuée aux adverbes, syntagmes prépositionnels et verbaux, propositions, (ou syntagmes nominaux équivalents) positionnés après le prédicat dans des phrases affirmatives (*e.g. Kim* ***phrases*** *the letter [with great care].*)

[7] *Obj : Object* - Tout complément d'objet du prédicat ou syntagme nominal dépendant du prédicat, positionné après lui et gouvernant le sujet d'une proposition complément du prédicat (*e.g. Voters* ***approved*** *the stadium measure. - They* ***expect*** *[us] to finish soon.*)

La troisième ressource est un corpus annoté de 150 000 phrases issues du BNC. Il permet d'illustrer les propriétés sémantiques et morphosyntaxiques des frames.

FrameNet est une ressource à la fois sémantique et morphosyntaxique qui contient des indications

intéressantes pour la résolution de rôles, d'autant plus que certains *rôles* sont sémantiquement typés. Aujourd'hui, elle compte plus de de 11 600 unités lexicales[13], réparties en 960 frames[14], ce qui correspond à 150 000 phrases annotées du BNC.

#### 2.2.1.4 Bilan

De nombreux efforts ont été fournis pour la constitution de ressources linguistiques sémantiques pour la langue anglaise. Ces efforts ont été nettement valorisés par l'usage courant des ressources telles que WordNet, Wikipédia, DBPedia ou FrameNet dans les tâches d'analyse sémantique. Ces ressources manuellement validées par les humains sont d'une précision très élevée.

L'unification de toute cette variété de ressources est un processus encore en cours. Tant que les différentes campagnes d'évaluation n'utiliseront pas de ressources unifiées, les recherches applicatives ne les emploieront pas non plus, n'ayant pas de banc d'essai à qui se comparer.

Excepté le cas des ressources collaboratives pour lesquelles les critiques suivantes sont à prendre avec quelque modération, l'inconvénient principal de ces ressources est leur non-exhaustivité. Il est très coûteux en temps et en ressources humaines de développer de telles bases lexicales de façon la plus complète possible.

De plus, ce mode de constitution ne permet de traiter qu'une langue à la fois. Ce qui a pour conséquence qu'en pratique, hormis les dictionnaires classiques, ces ressources ne sont quasiment développées que pour la langue anglaise.

Toujours concernant la non-exhaustivité de ces ressources, les langues évoluant en permanence, ce type de ressources ne peut avoir une couverture totale d'une langue à un instant $t$. Ceci est d'autant plus vrai pour les entités nommées : nous rappelons ici l'exemple assez drôle d'un système de reconnaissance vocale qui transcrivait *baraque aux Bahamas* pour un certain *Barack Obama*.

L'intérêt de ces ressources est indéniable pour leur qualité mais l'inconvénient principal reste le fait qu'elles ne sont développées qu'en très grande majorité pour la langue anglaise. Ceci est dû au coût de constitution que cela représente, ce coût étant tout aussi gênant pour leur maintenance et leur enrichissement.

### 2.2.2 Ressources automatiques

Nous venons de voir qu'un des inconvénients majeurs des ressources manuelles réside dans leur coût de constitution et leur faible couverture. Il existe par ailleurs une autre limite à ces ressources : elles ne contiennent généralement que peu d'information statistique.

13. nombre annoncé par le site officiel `http://framenet.icsi.berkeley.edu`, cependant on n'en trouve que 10196 dans la version téléchargeable FrameNet 1.3
14. 795 dans FrameNet 1.3

On pourrait par exemple désirer connaître la fréquence d'un mot dans une langue donnée, ou encore la fréquence du sens d'un mot par rapport à un autre sens de ce même mot, et ceci n'est pas possible avec l'usage d'un dictionnaire constitué par des experts humains (ou alors cette information est fondée sur des intuitions linguistiques). WordNet répond néanmoins à ce besoin en attribuant à un score de fréquence à chaque association mot synset. Ce score est calculé à partir du nombre d'occurrences du mot donné pour le synset donné sur un corpus désambiguïsé.

Pour aller plus loin, l'inventaire des sens fournis par une ressource manuelle suppose que la discrimination des sens d'un mot soit purement distincte. Or un certain nombre de travaux soutiennent que les sens et usages d'un mot sont des variations beaucoup plus complexes (chevauchement, sens n'apparaissant que dans un seul usage spécifique, compositionnalité, phraséologie, variation continue) et qu'il n'existe pas réellement de sens distincts si ce n'est pour une tâche donnée ([Atkins 1994], [Kilgarriff 1997], [Chibout 1998], [Vilnat 2005], [Mel'čuk 2010]). Aujourd'hui, une forte tendance consiste à caractériser cette variation continue par l'analyse de données en corpus.

Certains modélisent cette continuité des sens par des *graphes conceptuels*[15]. Dans ce cadre là, la stratégie manuelle adoptée par [Chibout 1998] consiste à formaliser les définitions de 2000 verbes en graphes conceptuels. Chaque graphe de définition d'un verbe se rattache à celui de son hyperonyme pour l'ensemble du sens qu'ils partagent, et introduit le ou les sème(s) spécifique(s) au verbe donné. Une telle approche permet non seulement un calcul de similarité entre verbes à l'intérieur du graphe, mais surtout une analyse extrêmement fine des métaphores (même si elles étaient inconnues jusque là), notamment par l'identification du sème dénominateur commun entre le verbe employé et sa signification métaphorique et par la détection rendue possible des analogies à l'origine des métaphores conceptuelles (cf. [Vilnat 2005]).

Une autre tendance consiste à induire cette fois-ci de façon automatique des clusters de sens dont la granularité peut varier en fonction des besoins applicatifs. Nous ne croyons pas en un modèle discontinu du sens comme le représente les dictionnaires traditionnels, mais plutôt en un modèle continu comme cela est avancé entre autres par [Victorri & Fuchs 1996], [Ploux & Victorri 1998], [Véronis 2004].

Nos propres travaux appartenant à cette deuxième tendance, nous présentons ici les diverses ressources automatiques portées à notre connaissance ainsi que leurs méthodes de constitution.

#### 2.2.2.1 Patrons lexicaux et acquisition de relations sémantiques

Parmi les travaux concernant la constitution de ressources automatiques, on trouve de nombreux travaux concernant l'acquisition de relations sémantiques à partir de larges corpora. Cependant, l'exercice étant assez proche de l'enrichissement d'ontologies, nous traiterons de ce sujet dans la section 2.3.2.1 qui leur est dédiée.

15. théorie développée par [Sowa 1984]

#### 2.2.2.2 Acquisition de lexiques syntaxiques

Un autre domaine de l'acquisition automatique concerne les cadres de sous-catégorisation. Nous n'avions jusque là mentionné l'existence de ces cadres uniquement lors de la description de FrameNet.

Les cadres de sous-catégorisation décrivent les différentes réalisations syntaxiques des arguments des verbes, ils définissent différentes combinaisons de structures et de fonctions grammaticales qui peuvent être utilisées avec un verbe donné. Ils donnent parfois des informations complémentaires sur les fréquences relatives des différents cadres, sur des préférences de sélection de traits sémantiques (animé/non animé, locatif/directionnel, etc), ou encore sur la nature lexicale de ces arguments,.

La constitution de telles ressources passe essentiellement par de l'analyse de corpus, qu'elle soit manuelle ou automatique. Cependant, dans le but d'atteindre la couverture la plus large possible et de constituer une ressource à large échelle regroupant tous les cas possibles, l'acquisition automatique devient incontournable. Un certain nombre de travaux, comme ceux de [Briscoe & Carroll 1997], [Korhonen 2002] et [Gardent & Lorenzo 2010] consistent à acquérir automatiquement des lexiques syntaxiques dans lesquels sont conservés à la fois les réalisations lexicales et syntaxiques des arguments.

#### 2.2.2.3 Les espaces sémantiques

Les espaces sémantiques, aussi appelés espaces distributionnels, espaces de mots ou en anglais *Word Space Model*, sont construits en exploitant l'information distributionnelle. Cette information distributionnelle permet de disposer les mots dans des espaces vectoriels de telle façon que la distance vectorielle séparant un mot d'un autre soit représentative de la différence de sens que l'humain perçoit entre ces deux mots. L'hypothèse de [Harris 1985] suggère que le sens d'un mot dépend de son contexte. C'est à partir de cette hypothèse que sont construits les espaces sémantiques, pour lesquels chaque dimension de l'espace correspond à un contexte précis et les valeurs associées pour un mot $x$ correspondent au nombre de fois où le mot $x$ est rencontré dans le contexte donné, dans un grand corpus représentatif d'une langue.

Beaucoup de paramètres entrent en jeu dans la constitution d'un espace sémantique et chacun de ces paramètres donne des caractéristiques différentes à l'espace obtenu. Dans les premiers travaux sur le sujet ([Salton *et al.* 1975]), les valeurs du Vector Space Model représentent la fréquence d'occurrence d'un terme donné dans chacun des documents (une colonne pour chaque document). Dans [Lund & Burgess 1996], [Ferret 2004], [Véronis 2004], les dimensions correspondent aux termes les plus fréquents du vocabulaire et les valeurs sont assignées au nombre de cooccurrences sur des fenêtres de taille fixe entre le terme de la ligne et le terme de la colonne (ou éventuellement l'information mutuelle). La phrase peut aussi être utilisée comme unité contextuelle, c'est le cas de [Ji *et al.* 2003]. Enfin, dans [Grefenstette 1994], [Strzalkowski 1994], [Lin 1998], [Padó & Lapata 2007],

[Grefenstette 2007], [Venant 2007] chaque dimension correspond à un contexte donné par l'intermédiaire d'une relation syntaxique donnée. [Curran 2004] compare également un certain nombre de contextes différents dont les contextes de fenêtre et les contextes issus de relations syntaxiques. Les meilleures mesures de similarité sont alors obtenues avec des contextes issus d'une analyse syntaxique de surface. Celle-ci permet d'analyser plus rapidement des gros corpus de texte qu'une analyse syntaxique profonde et donne des résultats équivalents voir meilleurs lorsque la taille du corpus disponible est suffisamment grande. Enfin on trouve dans les travaux de [Padó & Lapata 2007] un framework complet de consitution de tels espaces dans lequel le choix des contextes fait partie de l'ensemble des paramètres à déterminer avant de construire l'espace lui-même.

Parallèlement aux espaces sémantiques issus de corpus, on trouve aussi certains espaces sémantiques conçus à partir de ressources manuelles. C'est le cas des travaux de [Schwab *et al.* 2007] dans lesquels les vecteurs conceptuels représentent l'appartenance du terme donné à une liste de concepts variés donnée par un thésaurus (une colonne=un concept). On retrouve ce cas également dans les travaux de [Ploux & Victorri 1998] où les différentes dimensions sont données par le fait d'être synonyme ou non d'un autre mot.

**Analyse Sémantique Latente** La projection des mots dans des espaces vectoriels permit l'introduction d'une notion de distance spatiale fondée sur le sens entre les mots. Cette métaphore géométrique fut aussi la source d'une idée remarquable des psychologues [Landauer & Dumais 1997] qui utilisèrent les décompositions en valeurs singulières (cf. 2.1.2.1) pour découvrir la *structure latente* du vocabulaire que l'on utilise. Cette méthode appelée Analyse Sémantique Latente (*Latent Semantic Analysis, LSA*) fut sujette à de multiples variantes par la suite mais l'idée fondatrice reste la même. La matrice de cooccurrences est constituée à partir de contextes définis comme étant les documents. Elle est ensuite décomposée en valeurs singulières, puis reconstituée avec les $n$ premières valeurs singulières afin de ne conserver que le nombre de dimensions souhaitées. Les dimensions résultantes regroupent en quelque sorte en une seule des dimensions qui étaient corrélées entre elles dans la matrice d'origine. Ces dimensions étant des contextes d'occurrence, cela signifie que ces contextes étaient eux-mêmes assez corrélés et contenaient donc une part d'information redondante.

La matrice est donc réduite tout en conservant ses propriétés de proximité entre les différents éléments de l'espace et les auteurs montrent même une réduction du bruit qui améliore l'information contenue dans la matrice.

[Sahlgren 2006] dans son exploration plus théorique des modèles d'espaces de mots soutient que la technique LSA revient à simuler des espaces sémantiques paradigmatiques à partir d'espaces sémantiques syntagmatiques. D'après ses travaux et d'après la distinction syntagmatique/paradigmatique issue du *Cours de linguistique générale* de [Saussure (de) 1916] (sur laquelle se fondait également [Harris 1985]), les espaces construits à partir de cooccurrents sur des grands contextes sont plus à même de capturer des relations syntagmatiques entre les mots (mot apparaissant fréquemment dans les mêmes syntagmes), tandis que les espaces construits à partir de cooccurrents à un niveau plus

local (fenêtre, collocations) sont plus à même de modéliser des relations paradigmatiques (ou associatives) entre les mots. La LSA provenant du monde de la recherche d'information, les contextes de base sont des documents et capturent donc des relations syntagmatiques. En revanche la décomposition en valeurs singulières utilisée a pour effet de regrouper ensemble des mots qui ne cooccurrent pas nécessairement entre eux mais qui apparaissent dans des contextes similaires. C'est donc par ces biais que les espaces parviennent à la fin à capturer des relations paradigmatiques.

#### 2.2.2.4 Induction de sens de mots (*Word Sense Induction - WSI*)

D'après [Véronis 2004], se rendre indépendant des dictionnaires est primordial pour pouvoir atteindre une performance intéressante (notamment pour la Recherche d'Information). En utilisant les dictionnaires, même les annotateurs humains n'arrivent pas à concorder sur une désambiguïsation identique (Étude à l'aide du Petit Larousse : accord inter-annotateurs de 41% pour verbes et adjectifs, 46 % pour les noms). Les résultats obtenus sont similaires lorsque le référentiel de sens utilisé est WordNet. De ce fait, l'intérêt d'une induction automatique du sens des mots est grandissant, à la fois pour adopter une granularité de distinction plus adaptée aux usages, ainsi que pour faire face aux néologismes et aux spécificités du corpus étudié.

La tâche d'induction automatique de sens de mots est un problème appliqué de clustering. Il faut tout d'abord choisir quel sont les éléments à grouper en sous-ensembles représentant les sens du mot donné, et enfin choisir une stratégie de clustering.

Concernant les différentes sélections de termes à clusteriser, on trouve différentes variantes dans la littérature. [Ploux & Victorri 1998] rassemblent et mélangent d'abord plusieurs ensembles de synonymes issus de divers dictionnaires, certains distinguant des sens, d'autre non, et les clusterisent ensuite. Dans [Pantel & Lin 2002] ou [Velldal 2005], les clusters sont constitués en une seule fois à partir de tout le vocabulaire, chaque mot pouvant appartenir à plusieurs clusters. Ces clusters représentent donc des classes de synonymes et les mots appartenant à plusieurs clusters voient ainsi leurs sens discriminés. Les approches de [Yarowsky 1995] et [Schütze 1998] consistent à clusteriser pour chaque mot pour lesquels on veut induire des sens, toutes les instances de celui-ci, c'est-à-dire tous les contextes complets d'occurrence dans un corpus (phrase), discriminant ainsi leurs sens. L'approche de [Bordag 2006] est assez similaire dans le choix des éléments à clusteriser puisqu'il s'agit de clusteriser des triplets représentant des instances de mots ambigus en contexte. Enfin, [Dorow & Widdows 2003], [Rapp 2004], [Ji *et al.* 2003], [Véronis 2004], [Ferret 2004] clusterisent des cooccurrents des mots à distinguer en sens.

Au niveau des différentes méthodes de clustering employées pour l'induction de sens de mots, on trouve à la fois l'application d'algorithmes génériques ou la mise au point d'algorithmes spécifiquement adaptés à ce problème.

L'approche employée par [Ploux & Victorri 1998] et [Ji *et al.* 2003] consiste à détecter les cliques respectivement au sein des graphes obtenus par la modélisation des relations de synonymie et des graphes de cooccurrence.

L'approche de clustering *Shared Nearest Neighbors* est proposée par [Ertöz *et al.* 2001] pour la tâche de classification thématique de texte. Elle exploite l'idée que beaucoup de proches voisins partagés signifie une très forte similarité sémantique. Après avoir construit le graphe des plus proches voisins d'un mot $m$ , on construit le graphe des voisins partagés. Celui-ci prend pour valeurs sur ses arêtes le nombre de proches voisins partagés. Ensuite l'algorithme détermine des représentants de clusters (noyaux) que sont les noeuds de haut degré , auxquels seront adjoints les noeuds les plus fortement liés. Ces sous-graphes représentent alors les clusters de sens du mot $m$. Cette approche est reprise avec succès par [Ferret 2004] dans le clustering de cooccurrents pour l'induction de sens de mots.

[Pantel & Lin 2002] proposent l'algorithme *Clustering by comittee (CBC)*. La nouveauté de cette méthode de clustering réside dans le point suivant. Après une première phase de production de clusters potentiels appelés *comittees*, chaque mot de l'ensemble de départ est attribué itérativement à plusieurs de ces comités. Ceux-ci représenteront les différents sens du mot. A chaque fois que le mot est attribué à un des comités, on lui enlève les caractéristiques (traits) qui se chevauchent avec celles du cluster, avant de procéder à l'assignation suivante. Cela permet de découvrir des sens très distincts qui sont parfois beaucoup moins fréquemment utilisés et donc cachés par les sens fréquents.

L'article présente aussi une méthode d'évaluation automatique avec un mapping sur les sens de WordNet et une évaluation manuelle pour validation de la méthode automatique.

A la suite de ces travaux, [Tomuro *et al.* 2007] font l'observation suivante : l'algorithme CBC comme beaucoup d'autres, ne permet pas de distinguer des clusters contenant des mots polysémiques possédant une structure distributionnelle identique. Par exemple, on ne sait pas si un cluster contenant les mots *warm, cold* réfère au sens de la température physique ou du caractère d'une personne. Pour remédier à ce problème de clusters ambigus, ils proposent une mesure de similarité inter-traits permettant au score de similarité global entre deux vecteurs de tenir compte de la forte corrélation ou non des traits les plus saillants. Si deux vecteurs ont des valeurs fortes pour deux traits différents mais très fortement liés, alors cette similarité sera également prise en compte dans le score. Outre cette modification de score, les auteurs suggèrent aussi de ne conserver que les traits les plus saillants des centroïdes de comités, de sorte que si un comité était ambigu, on espère que seul son sens principal sera représenté par les traits les plus saillants, et qu'un autre comité pourra être constituer pour représenter le second sens. Ces modifications donnent de très bons résultats sur les adjectifs étudiés.

Parmi les autres travaux évoqués plus haut, on peut noter l'approche de [Rapp 2004] qui effectue une décomposition en valeurs singulières (SVD) suivie d'un clustering hiérarchique. D'après lui, la SVD lui permet de s'abstraire du problème d'insuffisance des données dû au choix des seuls 30 mots

les plus fortement liés.

#### 2.2.2.5 Bilan

En ce qui concerne l'induction de sens de mots, malgré l'étude comparative de [Purandare 2004], il n'existe pour l'instant aucune affirmation d'une certaine corrélation entre le type d'ambiguïté discriminée et les différents paramètres utilisées, tant dans la construction des espaces, que dans l'induction de sens même.

Nous n'avons pas évoqué dans cette section les méthodes employées pour évaluer les sens de mots induits mais nous reviendrons sur cette problématique au à la section 2.4.1.6 lorsque nous traiterons des méthodes de désambiguïsation utilisant des sens de mots induits.

Les méthodes d'acquisition automatique fournissent à la communauté du traitement automatique des langues un certain nombre de ressources complémentaires. Elles permettent entre autres d'extraire des relations sémantiques entre les mots, de donner des mesures de similarité et de dissimilarité sémantiques entre eux et enfin de fournir un inventaire continu des sens des mots. Ceci correspond au pendant des dictionnaires et ontologies constituées manuellement.

Malgré leur précision imparfaite, ces ressources automatiques sont indispensables pour pallier les problèmes de couverture non exhaustive et de coût de constitution de ressources manuelles. De plus, elles permettent de suivre l'évolution d'une langue ou l'apparition de nouvelles entités nommées non répertoriées par les ressources manuelles.

### 2.2.3 Conclusions

#### 2.2.3.1 Combiner et unifier les ressources

Bien que complémentaires, toutes les ressources répertoriées ici sont de natures différentes tant au point de vue sémantique que structurel, et ceci en omettant de discuter le formalisme de stockage. Beaucoup de tâches de traitement des langues et d'applications plus proches de l'utilisateur bénéficieraient d'une unification de ces ressources, essentiellement afin de profiter de leurs caractéristiques complémentaires.

#### 2.2.3.2 Normes et standards

Afin de pouvoir plus facilement et plus efficacement combiner l'ensemble des ressources lexicales numériques produites à l'heure actuelle, un format commun de stockage de ces données est nécessaire.

L'Organisation Internationale de Normalisation (ISO) se penche sur la question depuis 2001 par l'intermédiaire de son comité technique TC37 *Terminologie et autres ressources langagières et ressources de contenu*. Le sous-comité SC34 responsable de la *Gestion des ressources linguistiques* réunit 24 pays et un certain nombre d'organismes dont ELRA [16]. Presque 10 ans d'échanges et de

16. European Language Resources Association

révisions de son groupe de travail sur les *ressources lexicales* ont finalement abouti à un standard de description commun à toutes les ressources lexicographiques, le *Lexical Markup Framework* (LMF) ([Francopoulo *et al.* 2006]), issu de la norme *ISO 24613:2008* et dont la dernière version date de Mars 2008.

À l'heure actuelle, seul un petit nombre de projets et entreprises utilise ce standard de balisage lexical : ArabicLDB, BootStrep [17], Kyoto [18] (WordNet-LMF et Global WordNet Grid)), Lirics [19], Nedo, ProlexBase [20], et les entreprises Proxem, Spotter, et Tagmatica.

En revanche, les langages de représentation des connaissances publiés par le World Wide Web Consortium (*W3C*†) pour la constitution d'ontologies dans le cadre du *Web Sémantique*† sont de plus en plus utilisés par une communauté aux frontières du *TAL*. Le langages RDF (*Resource Description Framework*) est le premier à avoir été défini, puis il fut successivement étendu à RDF-S (*RDF Schema*) et OWL (*Web Ontology Language*), permettant ainsi de décrire des ontologies de plus en plus complexes.

La construction de base en RDF est le triplet {sujet, prédicat, objet}. Un langage d'interrogation associé à ce format de stockage a aussi été développé : le SPARQL. Un très grand nombre de ressources et d'applications du Web ont déjà une représentation sous forme de triplets RDF. On peut citer en autres : DBPedia (Mise sous forme de triplets d'un certain nombre de données entrées dans Wikipédia), Geonames, Flikr, Mozilla (pour ses marque-pages), Dublin Core (ressource bibliographique).

#### 2.2.3.3 Concrètement...

Parmi les avancées concrètes dans l'unification de ressources et l'usage de standards communs, on note les travaux de [Suchanek *et al.* 2007] dans lesquels les entrées de Wikipédia et les synsets de WordNet sont unifiés dans une ressource commune du nom de YAGO [21]. La ressource YAGO est également enrichie automatiquement par extraction automatique de relations et d'instances (cf. 2.2.2.1). Le modèle de données de cette ressource a été construit sur une extension du standard RDFS et est donc compatible avec d'autres ressources et algorithmes pouvant l'exploiter.

#### 2.2.3.4 Vers l'enrichissement de ressources

Nous avons vu qu'il était possible de constituer des ressources de façon automatique par l'analyse de grands corpus. Nous allons voir dans la section suivante comment, toujours de façon automatique, il est possible de traduire les ressources existantes (manuelles ou automatiques) et de les enrichir.

17. `http://www.bootstrep.org/bin/view/Extern/WebHome`
18. `http://www.kyoto-project.eu/`
19. `http://lirics.loria.fr/`
20. `http://www.cnrtl.fr/lexiques/prolex/`
21. `http://www.mpi-inf.mpg.de/yago-naga/yago/`

## 2.3 Des ressources existantes à leurs extensions

Nous nous intéressons principalement aux travaux de traduction et d'enrichissement des ressources WordNet et FrameNet car ce sont celles qui nous ont semblé combiner les avantages suivants :

- ces ressources disposent d'une bonne couverture de la langue anglaise et bénéficient de l'expertise d'excellents lexicographes ;
- ce sont des ressource libre d'accès pour la recherche ;
- une grande communauté les utilise et rend donc possible la comparaison avec des systèmes existants exploitant ces ressources.

Ce sont donc principalement ces ressources que nous utiliserons dans nos systèmes d'analyse sémantique.

### 2.3.1 La transcription de ressources manuelles vers d'autres langues

Les ressources linguistiques les plus massives sont souvent constituées par la communauté anglophone du fait de sa taille et des avancées scientifiques plus profondes sur le traitement de l'anglais comparé aux autres langues. Ceci boucle d'ailleurs un cercle vicieux puisque plus les ressources seront développées pour la langue anglaise, plus les systèmes de traitement automatique auront tendance a se développer et à être évalués autour de cette langue.

Il convient donc d'utiliser à bon escient les efforts déjà fournis et d'en tirer parti en exploitant les ressources manuelles anglophones dans le but de produire des ressources pour les autres langues. Nous nous intéressons plus particulièrement à la langue française mais sommes également intéressés par toute méthode permettant une telle transcription vers une autre langue. Une telle méthode pouvant soit être directement appliquée au cas du français si les deux langues sont suffisamment proches, ou du moins fournir quelques pistes supplémentaires pour la production de ressources francophones.

#### 2.3.1.1 Traduction de WordNet

Outre les travaux manuels de traduction du projet EuroWordNet (voir section 2.2.1.2), on recense un certain nombre de tentatives de constitution automatique de WordNets pour d'autres langues que l'anglais. Parmi celles-ci, [Sagot & Fišer 2008] distingue deux grandes tendances : les approches par fusion (*merge approach*) parmi lesquelles se situent les approches de [Kotis *et al.* 2006] et [Falk *et al.* 2009] et les approches par extension (*expansion approach*) comme celles proposées par [Sagot & Fišer 2008] ou [Barbu & Barbu Mititelu 2005].

Les approches par fusion consistent à construire des ontologies indépendamment et de déterminer un mapping avec les WordNet existants *a posteriori*. L'avantage d'une telle approche est que l'on peut s'abstraire de la structure existante de WordNet. Dans un cas où une langue n'a pas une richesse identique à l'anglais pour certaines parties du réseau lexical, cela permet de constituer une hiérarchie plus adaptée à la langue.

Dans le cas du français, on peut supposer dans un premier temps que la distribution est suffisamment proche de celle de l'anglais pour constituer la ressource à partir du WordNet original et de l'étendre ensuite en cas de nécessité. C'est ce que l'on appelle approche par *extension*.

Pour traiter le problème de la polysémie, beaucoup de travaux utilisent un dictionnaire bilingue pour y sélectionner les traductions les plus pertinentes selon diverses heuristiques. C'est le cas de [Barbu & Barbu Mititelu 2005]. Une approche originale de [Sagot & Fišer 2008] utilise des corpus parallèles. Chaque langue possédant un WordNet permet d'annoter la partie de corpus correspondante sans désambiguïsation : tous les synsets correspondant à un mot sont utilisés pour l'annotation. Les identifiants de synsets étant les mêmes pour chaque langue, et les différentes traductions d'un même mot ne conservant pas les mêmes sens dans chaque langue, le système peut alors désambiguïser les termes français par calcul de l'intersection des annotations des termes parallèles de chaque langue. Chaque terme annoté en français dans le corpus est ensuite injecté dans la nouvelle ressource. La ressource française ainsi construite, Wolf (WordNet pour la Langue Française) contient environ 32 000 synsets.

L'approche de [Falk *et al.* 2009] pour la constitution de groupes de synonymes discriminés par sens est une approche de constitution de WordNet par fusion. Il ne s'agit pas d'une traduction de ressources anglophones existantes. Les auteurs partent de ressources francophones, le dictionnaire TLFi (Trésor de la Langue Française informatisé) pour la distinction en sens ainsi que 5 dictionnaires de synonymes, et essaient de grouper les synonymes des mots cibles en fonction des différents sens du mot cible dans le TLFi. La F-mesure obtenue sur les verbes, catégorie relativement difficile à désambiguïser est de 0,7. Les auteurs mentionnent leur intention de lier ces groupes de synonymes aux WordNets français existants.

#### 2.3.1.2 Traduction de FrameNet

Pour la langue française, les ressources utilisables dans le cadre d'une méthode *Semantic Role Labeling* sont plutôt rares. La seule ressource qui nous soit connue est Volem [Fernandez *et al.* 2002]. Il s'agit d'une base en langues française, espagnole et catalane regroupant 1 500 verbes pour lesquels un petit nombre de rôles communs sont définis. Cependant cette ressource lexicale reste d'une taille négligeable comparée aux ressources constituées pour la langue anglaise. De plus aucun corpus n'y est rattaché et seuls les exemples (non annotés) peuvent être utilisés comme données d'apprentissage.

Pour subvenir aux besoin des systèmes de langue non anglophone, plusieurs méthodes ont été employées pour traduire des ressources de *SRL* dans différentes langues. La méthode de traduction de FrameNet proposée par S. Padó et M. Lapata [Padó & Lapata 2005] consiste à annoter automatiquement la partie anglaise d'un corpus parallèle aligné de façon intraphrastique. Par projection sur la langue cible ils associent ensuite chaque *frame* à un ensemble d'unités lexicales déclencheuses et chaque rôle à un ensemble de syntagmes. Cette méthode a l'avantage de produire d'une part une base de données FrameNet pour la langue cible et d'autre part un corpus d'apprentissage à l'usage

d'un annotateur automatique.

Des traductions ont ainsi été obtenues pour l'allemand [Padó & Lapata 2005] , l'italien [Tonelli & Pianta 2008] et le français [Padó & Pitel 2007].

[Basili *et al.* 2009] proposent une approche également fondée sur des corpus parallèles mais elle se distingue des précédentes par l'avantage de ne pas nécessiter d'analyseur syntaxique. L'appariement des prédicats à leur frame utilise des distances issues d'un espace sémantique après avoir calculé un vecteur représentatif de la phrase donnée dans ce même espace. Le transfert des frames se fait selon différentes heuristiques exploitant les probabilités de traduction d'un traducteur automatique. Les résultats bien que non formellement comparés à ceux de [Tonelli & Pianta 2008], semblent s'avérer meilleurs.

Les résultats de ces deux derniers travaux fournissent des corpus d'apprentissage pour la langue italienne.

La seule autre ressource française de type FrameNet que nous connaissons est issue du projet Luna [22]. Il s'agit d'un petit sous-ensemble de frames (cf. Figure 2.3) construites manuellement et indépendamment du FrameNet de Berkeley par [Meurs *et al.* 2008]. Le but était de s'adapter le plus possible à la fois au domaine spécifique (réservation de chambre d'hôtel) et à la modalité orale du corpus MEDIA.

Bien que ce ne soit pas la traduction d'une ressource existante, nous mentionnons également les travaux de [Falk & Gardent 2010] sur la constitution de classes de verbes français en fonction de leurs cadres de sous-catégorisation. La ressource générée est une ressource assez similaire à VerbNet contenant la classification de 3 536 verbes (3 626 dans VerbNet), distingués en 500 classes. Chaque classe est définie par un ensemble d'attributs syntaxiques et sémantiques que tous ses éléments partagent. D'après [Levin 1993], une telle classification permet aux verbes appartenant à une même classe de partager un ou plusieurs aspects sémantiques.

#### 2.3.1.3 Bilan

Malgré ces différents travaux sur la constitution de ressources pour la langue française, celles-ci restent encore d'une qualité et d'une taille assez faibles. Dans l'optique d'une exploitation de telles ressources pour l'analyse sémantique de textes français, notre travail se concentrera sur la constitution ou l'amélioration des ressources existantes.

D'autre part ces ressources n'étant pas exhaustives (comme vu précédemment, du fait du problème initial de couverture et des problématiques liées aux néologismes et aux glissements de langue), nous nous intéressons aussi aux problématiques d'enrichissement automatique.

22. http://www.ist-luna.eu/index.htm

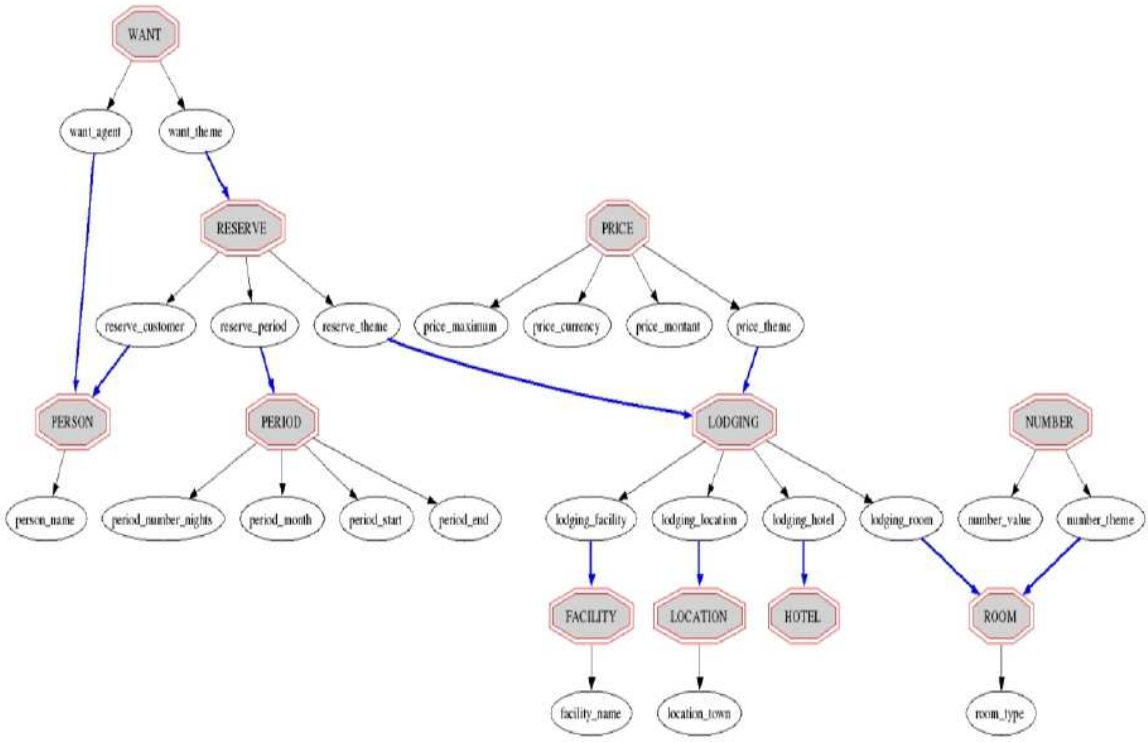



FIGURE 2.3 – Frames définies dans le projet Luna. Source : [Meurs *et al.* 2008]

### 2.3.2 L'enrichissement automatique de ressources manuelles

Nous nous intéressons maintenant aux travaux de la littérature traitant le problème de la non-exhaustivité des ressources manuelles WordNet et FrameNet.

En effet, dans WordNet, les instances de classe sont souvent liées à l'actualité et à la culture générale du monde anglophone. Il manque donc une quasi-infinité de nouvelles entités nommées ou d'entités liées à la culture d'autres pays. De plus, on peut trouver des relations sémantiques encore non répertoriées dans WordNet.

Dans le cas de FrameNet, certaines unités lexicales devraient figurer dans la ressource mais n'appartiennent à aucune cadre (e.g. *taxonomy.n* devrait figurer parmi les unités lexicales de *Categorization*), tandis que d'autres y figurent mais n'appartiennent pas à toutes les cadres qu'elles devraient déclencher (e.g. *boom.n* n'apparaît qu'en tant que *Sounds* alors qu'on aimerait le voir aussi dans *Progress* comme dans l'expression *en plein boom*).

#### 2.3.2.1 Enrichissement d'ontologies

Afin d'enrichir les ontologies existantes, il est nécessaire de caractériser des propriétés que les relations partagent. Ainsi, il devient possible de repérer de telles relations dans des corpora et d'insérer les mots concernés dans les ontologies. On peut distinguer différentes approches en fonction du type de relations recherchées.

Les relations taxonomiques ou purement sémantiques sont des relations concernant les sens relatifs de termes, cela comprend par exemple les synonymes, les antonymes, les hyperonymes/hyponymes ou les holonymes/méronymes. La ressource de référence utilisée pour comparaison ou enrichissement est très souvent WordNet.

Les travaux initiaux de [Hearst 1992] concernant l'acquisition automatique d'hyponymes à partir de patrons lexico-syntaxiques ont ouvert la voie à de nombreux travaux d'acquisition exploitant la redondance d'information que l'on trouve dans de larges corpora (et sur le Web en particulier). Par la suite, les approches exploitant les patrons d'extraction pour l'acquisition de relations proposèrent d'apprendre ces patrons automatiquement. Pour un état de l'art plus complet sur l'extraction de relations fondée sur les patrons lexico-syntaxiques, le lecteur peut se référer à [Auger & Barrière 2008].

Par ailleurs, [Chalendar (de) 2001] propose un système complet d'acquisition automatique de ce type de relations sur des domaines variés. L'idée principale est de regrouper tous les noms apparaissant dans un même contexte défini par un même prédicat et une même relation syntaxique, lorsque les segments de texte correspondants sont préalablement classifiés comme appartenant à une même thématique. Certains comme [Ruiz-Casado *et al.* 2006] s'intéressent à l'enrichissement de WordNet par l'acquisition de relations sur les textes de Wikipédia. [Snow 2006] propose un algorithme probabiliste pour la classification de relations d'hyperonymie et de termes cousins et l'introduction de

nouveaux termes et relations dans WordNet. Enfin, [Morin & Jacquemin 2004] décrivent un framework complet d'acquisition de termes en relations d'hyponymie (incluant les *expressions multimots*[†]). Ils présentent une méthode pour apprendre de nouveaux patrons en fonction de termes sources fournis au système. Ils proposent ensuite d'exploiter les variations syntaxiques, morphosyntaxiques et sémantiques des termes multimots pour étendre les liens appris sur les termes simples à des termes multimots.

Il existe également un nombre important de travaux sur la détection des mots de sens proches dans les espaces distributionnels. C'est le cas des travaux de [Grefenstette 1994], [Landauer & Dumais 1997], [Turney 2001], [Van Der Plas & Tiedemann 2006], [Ferret 2010]. Les travaux de thèse de [Curran 2004] fournissent une étude très complète sur le mode d'évaluation et les différents paramètres d'obtention d'une bonne mesure de similarité sémantique : contextes à utiliser, poids à appliquer au compte d'occurrences (fréquences, MI, $\chi^2$, t-test, Dice, Tf-Idf...), mesure de similarité (distance géométrique, cosinus, Jaccard, [Grefenstette 1994]...), stratégie d'extraction de synonymes. Certains travaux définissent également des mesures de distances calculées à partir de ressources manuelles telles que WordNet ([Patwardhan *et al.* 2003], [Pedersen *et al.* 2004]).

Les relations sémantiques événementielles regroupent des relations telles que *date de naissance, capitale, rachat, mariage, descendance, direction, casting...* Parmi ces systèmes, Snowball ([Agichtein & Gravano 2000]) appartient aux premiers systèmes d'extraction de relations. [M. Suchanek *et al.* 2006] proposent un apprentissage fondé sur une analyse linguistique et statistique et rapportent une précision et un rappel bien supérieurs à ceux de ses prédécesseurs Snowball et Text2Onto ([Cimiano & Völker 2005]). [Pantel & Pennacchiotti 2006] présentent aussi de meilleurs résultats que leurs prédécesseurs mais les systèmes et corpus comparés ne sont pas les mêmes que dans [M. Suchanek *et al.* 2006]. Ce domaine est plus proche de l'extraction d'information.

Les relations sémantiques spécifiques concernent l'acquisition de relations spécifiques à un domaine d'expertise. Des études comme celles de [Embarek & Ferret 2008] montrent que ces méthodes sont également adaptables à des domaines métiers spécifiques comme le domaine complexe de la médecine.

En ce qui concerne l'instanciation de classes, on trouve par exemple dans le cadre du projet KnowItAll ([Etzioni *et al.* 2004]) une méthode probabiliste de peuplement d'ontologies où les classes et relations sont prédéfinies et où les patrons d'extraction et les critères de validation sont appris au fur et à mesure en fonction des instances apprises.

#### 2.3.2.2 Enrichissement de FrameNet

Plusieurs approches ont été proposées ces dernières années pour pallier le problème de couverture insuffisante de FrameNet freinant son utilisation dans des tâches applicatives.

[Johansson & Nugues 2007b] étudient deux approches utilisant WordNet pour étendre la couverture de FrameNet. Dans un premier temps, ils utilisent la similarité de WordNet appelée Lesk et définie par [Pedersen *et al.* 2004] pour calculer un score de similarité entre toute nouvelle unité lexicale et les unités lexicales déjà présentes dans les frames. La similarité d'une unité lexicale avec une frame est donnée par la moyenne des scores de similarité calculés avec chacune des unités lexicales présentes dans la frame donnée. L'unité lexicale est attribuée à la frame ayant le plus grand score de similarité. Dans un deuxième temps, les auteurs proposent d'entraîner un classifieur SVM sur un ensemble d'unités lexicales appartenant ou non aux frames. Les vecteurs caractéristiques des unités lexicales sont constitués de tous les synsets appartenant à l'ensemble des hyperonymes de tous les sens WordNet de l'unité lexicale donnée, pondérés par la fréquence relative d'occurrence de ces sens dans SemCor. Les meilleures résultats sont produits à l'aide de la deuxième méthode générant ainsi une ressource contenant 18 372 unités lexicales supplémentaires avec une précision estimée à 70 % (calculée sur 100 mots pris aléatoirement dans le FrameNet étendu).

[Pennacchiotti *et al.* 2008] proposent deux méthodes principales pour l'assignation de nouvelles unités lexicales aux frames qui leur correspondent le mieux. Les deux méthodes proposent de projeter à la fois la nouvelle unité lexicale et l'ensemble des frames de FrameNet dans un même espace et d'assigner la nouvelle unité lexicale à la frame dont elle est la plus proche dans l'espace défini.

La première méthode exploite des similarités fournies par des espaces sémantiques (3 essais différents : cooccurrences de fenêtre, cooccurrences syntaxiques et cooccurrence sur un chemin syntaxique de distance n). Dans ces espaces, chaque frame est représentée par le barycentre de leurs unités lexicales (chaque unité étant pondérée par un coefficient lié à sa fréquence dans un corpus externe). Chaque nouvelle unité lexicale est assignée à la frame dont elle est la plus proche.

La seconde méthode définit des sous-graphes de WordNet pour chaque frame et calcule des *densités conceptuelles* ([Agirre & Rigau 1996]) pour les différents éléments saillants des sous-graphes. Les nouvelles unités lexicales sont intégrées aux sous-graphes de toutes les frames et les densités conceptuelles sont recalculées. Chaque unité lexicale est assignée à la frame dont le sous-graphe possède le noeud avec la meilleure densité conceptuelle.

Si l'on ne considère que les frames avec le meilleur score, les résultats montrent une précision de 52% pour l'approche fondée sur WordNet et une précision de 27% avec l'approche distributionnelle. En revanche, l'approche distributionnelle est un bon *back-off* pour les cas où l'approche WordNet ne retourne pas de réponse. En effet, une approche combinée de cette façon apporte un gain en couverture de 15% pour une perte en précision de seulement 4% par rapport à l'approche WordNet correspondante utilisée seule.

#### 2.3.2.3 Bilan

L'enrichissement automatique des ressources lexicales manuelles existantes permet de résoudre en partie les problèmes de couverture de ces ressources. En effet, on a longtemps reproché à FrameNet de n'être pas assez exhaustif pour pouvoir servir de référence pour l'annotation des rôles. Des techniques d'enrichissement automatique de celui-ci apparaissent pour ajouter de nouvelles unités lexicales déclencheuses de frames à la ressource initiale. Les résultats en précision de ces méthodes varient encore entre 50 et 70% et peuvent encore être améliorés.

### 2.3.3 Conclusions

L'usage de méthodes automatiques pour la traduction et l'enrichissement des ressources existantes semble prometteur. Cependant, les ressources obtenues par ces méthodes restent à l'heure actuelle assez peu nombreuses pour le français et d'une précision encore insuffisante. Dans cette thèse, nous avons donc conçu et évalué de nouvelles méthodes pour la création de ressources de langue française. Celles-ci seront présentées au chapitre 3 à la section 3.1 pour la traduction automatique de WordNet et au chapitre 5 pour la traduction et l'enrichissement automatiques de FrameNet.

## 2.4 Tâches d'analyse sémantique

Nous nous intéressons aux deux tâches d'analyse sémantique présentées dans l'introduction. Nous présentons dans un premier temps les travaux menés sur la désambiguïsation lexicale et dans un deuxième temps les travaux menés en annotation sémantique de rôles depuis l'identification de ces deux problématiques jusqu'à nos jours.

### 2.4.1 La désambiguïsation lexicale

Une des difficultés majeures du traitement automatique des langues est la résolution des ambiguïtés lexicales. En effet, la plupart des mots que nous employons ne réfèrent pas à une unique signification mais ont plusieurs usages différents qui sont entre autres répertoriés par les dictionnaires (cf. sections 1.2 et 2.2.1).

Cette résolution est effectuée par le cerveau humain très naturellement plusieurs fois par phrase sans même que l'on ne s'en rende compte. Pour les machines, le problème n'est toujours pas considéré comme entièrement résolu après 60 ans de recherche sur le sujet.

Nous présentons ici tout d'abord l'évolution globale des différentes approches de désambiguïsation (*Word Sense Disambiguation - WSD*), et passons ensuite dans une revue plus détaillée les systèmes qui nous paraissent les plus essentiels parmi les différentes approches.

Dans leur éditorial du numéro dédié au *Word Sense Disambiguation* du journal *Computer Speech and Language*, [Preiss & Stevenson 2004] évoquent plusieurs problématiques liées à cette tâche : de quels lexiques la communauté dispose-t-elle pour répertorier les sens des mots à désambiguïser ? Comment décider si l'annotation automatique doit attribuer un ou plusieurs sens aux mots ? Certains proposent d'attribuer plusieurs sens possibles, du fait des accords inter-annotateurs très variables ([Ahlswede & Lorand 1993], [Kilgarriff & Rosenzweig 2000]), la granularité des lexiques utilisés jouant un rôle important dans la valeur de ces taux d'accords. En ce qui concerne l'évaluation [Preiss & Stevenson 2004] soulèvent le problème du framework proposé par la campagne d'évaluation Senseval qui n'est pas adapté à tous les systèmes. De plus, si toute la communauté utilise les mêmes ensembles de test, les performances risquent d'être influencées en fonction des phénomènes linguistiques figurant dans les données de la campagne. Les auteurs mettent donc en avant ce que [Véronis 2004] préconise, à savoir que la meilleure évaluation possible reste l'amélioration ou non d'une application finale exploitant le système (dans son cas, il s'agit d'une application de recherche d'information).

Cet éditorial présente aussi une synthèse des différentes tendances en WSD. Les systèmes évalués par rapport à une vérité terrain utilisent maintenant des lexiques à grande échelle contrairement à ce qui se faisait jusque là et où les systèmes n'étaient alors capables de ne désambiguïser qu'un nombre restreint de mots. [Preiss & Stevenson 2004] remarquent aussi que la disponibilité d'un certain nombre de lexiques a permis la mise en œuvre de différentes stratégies de désambiguïsation exploitant chacune des informations différentes, présentes dans certaines ressources et pas dans d'autres. Par ailleurs, beaucoup d'approches utilisent des corpus, qu'ils soient internes au framework d'évaluation ou complètement externes. Ces corpus peuvent être annotés par sens, par d'autres connaissances, ou même non annotés ou enfin il peut également s'agir de corpus parallèles alignés. En outre, la combinaisons de différentes ressources et/ou de différents algorithmes de désambiguïsation se fait de plus en plus ([Ng & Lee 1996], [Hirst 1987], [Stevenson & Wilks 2001]). Enfin, les techniques d'apprentissage automatique commencent à être régulièrement utilisées et à produire de bons résultats ([Véronis & Ide 1990], [Yarowsky 1993]).

Dans un état de l'art plus récent, [Agirre & Edmonds 2007] dressent un historique des travaux sur la question. Nous reprenons ici les éléments qui nous ont semblé les plus importants dans l'évolution des systèmes de WSD. D'après ces auteurs, la première évocation du problème de désambiguïsation lexicale dans la littérature scientifique date de la fin des années 40, lorsque la recherche en traduction automatique fait ses premiers pas. [Weaver 1949] met le problème en évidence dans l'introduction de son memorandum sur la traduction automatique :

> If one examines the words in a book, one at a time through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of words. "Fast" may mean "rapid"; or it may mean "motionless"; and there is no way of telling which. But, if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N

words on either side, then, if N is large enough one can unambiguously decide the meaning...[23]

La même année, [Zipf 1949] publie une étude dans laquelle il présente ce qu'il nomme *Law of Meaning*[24]. En se fondant sur des ordres de fréquence sur les 20 000 mots les plus fréquents de la langue anglaise généraliste (rapportés par [Thorndike 1941]), cette étude montre la corrélation entre la fréquence des mots et leur nombre de sens. Plus ils sont fréquent et plus ils sont polysémiques. Ceux-ci suivent en effet une loi de puissance. Cette relation fut également vérifiée plus tard sur le British National Corpus par [Edmonds 2005].

L'intérêt des travaux en traduction automatique pour la désambiguïsation lexicale se maintient par la suite et les premières tentatives de [Masterman 1957] utilisant les catégories du *Roget's International Thesaurus* donnent déjà l'orientation des travaux à venir sur la désambiguïsation, à savoir notamment le besoin d'une ressource linguistique de référence. En revanche, devant la difficulté de la tâche, les travaux de traduction automatique sont quasiment abandonnés dans les années 60.

A la fin des années 70, [Rieger & Small 1979] soumettent l'idée des *word experts*, des processus indépendants liés à chaque mot dont la tâche est la désambiguïsation. [Hirst 1987] développe cette idée avec les *Polaroid Words* qui permettent à la fois de désambiguïser lexicalement les mots polysémiques et d'annoter le texte en *frames* et rôles suivant la base de connaissance de [Wong 1982] (base similaire à FrameNet mais de taille beaucoup plus restreinte).

Les années 80 voient l'arrivée des corpus à grande échelle, l'acquisition automatique devient alors possible ([Wilks *et al.* 1990]). [Yarowsky 1992] combine alors l'information du thésaurus Roget avec des informations de cooccurrences issues de l'analyse de grands corpus pour produire des règles de désambiguïsation sur les sens du thésaurus Roget.

Bien que les méthodes utilisant des dictionnaires soient intéressantes pour certains cas d'ambiguïté, leur robustesse reste limitée car elles souffrent d'un problème de couverture sur la distinction des sens. On assiste alors dans les années 90 à trois événements majeurs :

- WordNet [Miller *et al.* 1990] est constitué et mis à la disposition de tous. Cela constitue un grand pas en avant du fait de sa structure hiérarchique et de son accessibilité numérique. WordNet est aujourd'hui la ressource généraliste de sens la plus utilisée en recherche sur la désambiguïsation lexicale.

23. Si on observe les mots d'un livre un à un à travers un masque opaque ne laissant apparaître qu'un seul mot, alors il est évidemment impossible de déterminer, un par un, le sens des mots. *Fast* peut signifier *rapide* ; ou cela peut aussi bien vouloir dire *immobile*, *attaché* ; et il n'y a aucun moyen de trancher. Cependant, si on agrandit la fente du masque opaque jusqu'à ce qu'on ne voit plus seulement le mot en question mais disons aussi N mots de chaque côté, alors, si N est suffisamment grand, on peut décider de la signification sans qu'il ne reste aucune ambiguïté...

24. Loi du sens

- Les techniques statistiques font leur entrée dans la communauté du Traitement Automatique des Langues (apprentissage automatique supervisé ou non).
- La première campagne d'évaluation SENSEVAL a lieu. Les systèmes jusqu'à ce jour n'étaient que très peu comparables entre eux car il n'existait pas de benchmark commun. Toutes les évaluations menées jusque là différaient tant dans la variété des mots évalués, des annotateurs, des inventaires de sens et des corpus utilisés.

Enfin, [Gale *et al.* 1992a] fixent pour la première fois des limites inférieures et supérieures aux résultats d'évaluation par l'utilisation d'une baseline (choix du mot le plus fréquent) pour la limite inférieure et d'un accord inter-annotateurs pour la limité supérieure.

Nous dressons dans cette section l'inventaire des principales techniques de désambiguïsation lexicales trouvées dans la littérature. Nous décrivons dans un premier temps les approches de types Lesk. Nous passons ensuite en revue les systèmes de désambiguïsation par apprentissage automatique. Nous étudions tout d'abord les systèmes supervisés, puis les systèmes non supervisés et enfin une méthode utilisant le bootstrapping. Enfin, nous nous intéressons aux méthodes de désambiguïsation spécifiques aux sens induits ainsi qu'à leur évaluation.

#### 2.4.1.1 Algorithmes de type Lesk

[Lesk 1986] propose une approche exploitant les définitions apportées par les dictionnaires numériques (*Machine Readable Dictionaries, MRD*) comme connaissance externe aux phrases à traiter. En effet, les dictionnaires répertorient les différents sens des mots avec une définition spécifique à chacun d'entre eux. Le sens d'un mot ambigu est choisi pour la définition qui comptabilise le plus de mots en commun avec les définitions des mots qui se trouvent dans le contexte du mot ambigu dans la phrase. Le contexte est alors défini dans cette première expérience comme étant une fenêtre de taille fixe de 10 mots pleins autour du mot ambigu. Les précisions obtenues dans cette étude préliminaire varient entre 50 et 70%.

Cette approche sera reprise par la suite par un grand nombre d'études complémentaires. C'est le cas entre autres de [Vasilescu & Langlais 2004] qui étudient les performances de différentes variations de l'algorithme Lesk. L'idée est de prendre le *contexte* du mot à désambiguïser et de comparer les *descriptions* des mots de ce contexte avec un ensemble de *descriptions* des sens candidats. Le sens qui obtient le meilleur *score* est attribué au mot cible. Les auteurs utilisent plusieurs définitions du contexte, de la description et du score.

L'évaluation est ensuite menée sur le corpus Senseval-2 (2473 mots-cibles). Les résultats obtenus donnent lieu à divers conclusions sur le contexte optimal et la meilleure stratégie à adopter. En ce qui concerne le contexte, le passage d'un simple contexte de fenêtre symétrique de mots pleins à un contexte plus complexe filtrant les mots en fonction de leur force d'association au mot cible et intégrant additionnellement une *chaîne lexicale* liée à chaque mot retenu, apporte beaucoup, tant

en précision qu'en rappel. La chaîne lexicale du mot cible est définie d'après [Hirst & St Onge 1998]. Chaque mot plein apparaissant dans un contexte fenêtre autour du mot cible et ayant une mesure de similarité avec celui-ci supérieure à un certain seuil appartient à la chaîne lexicale. La mesure de similarité est définie comme étant l'*indice de Jaccard*† entre les ensembles de synonymes et d'hyperonymes (jusqu'au plus élevé) de chacune des deux entités.

La taille optimale du contexte est en général de 3 (à l'exception de l'algorithme de désambiguïsation Bayes pour lequel la taille optimale est de 25). D'autre part, la meilleure description en petit contexte semble être l'ensemble des mots pleins lemmatisés issus des définitions de WordNet tandis que sur des grands contextes, il apparaît plus intéressant de considérer l'ensemble des mots obtenus en suivant les liens d'hyperonymie et de synonymie de WordNet.

Concernant le choix de la meilleure stratégie, la stratégie utilisant la chaîne lexicale est dite *silencieuse*, les rares réponses différant de la réponse baseline sont prises à bon escient. Les risques négatifs sont donc faibles. Après une étude complémentaire, il apparaît que le filtrage morphosyntaxique (seule les définitions de la catégorie grammaticale identifiée sont conservées comme candidates) améliore les performances. Les auteurs remarquent également qu'avec un étiquetage morphosyntaxique (manuel ou automatique), la performance du meilleur algorithme ([SLESK + CL]) s'améliore effectivement. En revanche, sans la prise en compte de l'étiquetage morphosyntaxique, la performance (précision et rappel) de la stratégie proposée est meilleure que celle de l'algorithme *baseline* mais avec cette prise en compte, elle reste équivalente à la performance de l'algorithme *baseline*. La performance atteinte avec un étiquetage automatique est de 60,5% en précision et de 59,9% en rappel.

Les meilleurs résultats obtenus sur ces mêmes données dans la tâche *English All words* de la campagne d'évaluation Senseval sont de 60,9% en précision et rappel pour les systèmes supervisés et de 57.5% en précision et 56,9% en rappel pour les systèmes non supervisés. Les résultats produits par ce système non supervisé sont donc supérieurs à ceux précédemment obtenus. Cependant, la conclusion à laquelle les auteurs arrivent est que, comme l'observe [Véronis 2001], l'information contenue dans une ressource telle que WordNet n'est pas suffisante à une bonne désambiguïsation et que l'ajout d'information pragmatique et/ou syntaxique semble nécessaire.

#### 2.4.1.2 Systèmes d'apprentissage supervisé

Les méthodes d'apprentissage automatique ont été appliquées avec succès au problème de classification de sens, autrement dit à la désambiguïsation elle-même. Les approches par apprentissage sont devenues les plus courantes et ce sont celles qui permettent aujourd'hui d'obtenir les meilleurs résultats dans les campagnes Senseval.

Un des premiers systèmes de désambiguïsation utilisant un apprentissage supervisé est proposé par [Brown *et al.* 1991]. Cet algorithme est plus spécifiquement conçu pour faire de la désambiguïsation

de termes de traduction. [Brown *et al.* 1991] utilisent l'algorithme flip-flop ([Nadas *et al.* 1991]) pour produire deux classes de discrimination pour chaque mot ambigu. Chacune de ces deux classes produites possède des règles de discrimination ainsi que son propre ensemble de probabilités de traduction. Cette méthode permet le découpage le plus informatif possible entre les deux traductions les plus fréquentes, mais il reste une ambiguïté sur les traductions possibles d'une même classe.

[Gale *et al.* 1992a] proposent une approche bayésienne exploitant une règle de décision fondée sur les probabilités d'occurrences d'un contexte sachant le sens du mot cible dans le corpus d'apprentissage. Les résultats montrent qu'un contexte de grande taille donne de meilleurs résultats ($\pm$50 mots de chaque côté du mot cible).

[Yarowsky 1992] reprend la méthodologie employée par [Gale *et al.* 1992a] en injectant une notion supplémentaire de saillance de mots pour chaque catégorie du thésaurus Roget. Cette technique est particulièrement intéressante lorsqu'il s'agit de désambiguïser un texte dans un domaine précis. Ainsi, préalablement à la désambiguïsation, l'algorithme apprend sur un corpus du domaine (non annoté). Il attribue à chaque catégorie du thésaurus une liste de mots représentatifs associés à un score de saillance. Les mots représentatifs sont choisis en corpus lorsqu'ils apparaissent dans le contexte d'un mot appartenant à la catégorie dans le thésaurus. Ainsi lors de la désambiguïsation, il ne suffit pas que le mot d'un contexte appartienne à une catégorie du thésaurus pour donner plus de poids à celle-ci dans la décision. Le mot du contexte peut simplement apparaître fréquemment au contact de mots de cette catégorie. Les résultats obtenus sont bons sur des mots dont les différents sens sont bien dépendants du domaine, mais ce n'est pas le cas pour un grand nombre de mots, comme par exemple le mot *intérêt* (personnel) qui relève de quasiment tous les domaines.

D'autres adoptent encore d'autres stratégies, comme [Segond *et al.* 1997] qui entraînent un modèle de Markov caché sur les bigrammes de tags sémantiques issus de WordNet 1.5., [Crestan *et al.* 2003] qui utilisent des arbres bayésiens ou encore [Lee & Ng 2002] qui étudient le choix des meilleurs traits et des meilleurs algorithmes sur les données de Senseval 1 et 2 (système à base de *Support Vector Machine (SVM)*, Bayes naïf, AdaBoost, et arbres de décision).

Quinze ans après les débuts des méthodes statistiques en désambiguïsation, les méthodes supervisées donnent toujours les meilleurs résultats d'après les dernières évaluation Senseval et SemEval. Le meilleur système pour la tâche *all words coarse-grained* (et deuxième système pour la tâche *fine-grained*) de SemEval 2007 ([Chan *et al.* 2007]) utilise un apprentissage de type SVM sur le corpus SemCor ainsi que sur le corpus DSO [Ng & Lee 1996] et des textes bilingues chinois-anglais alignés (étayant ainsi l'hypothèse suivant laquelle plus de données d'apprentissage améliorent la performance des systèmes de désambiguïsation). A titre indicatif, la précision obtenue par ce système est de 82,5% pour la tâche *coarse-grained* et de 58,7% pour la tâche *fine-grained*, toutes les occurrences de termes ambigus étant traitées.

On retrouve également des modèles à Maximum d'Entropie comme dans les travaux de [Tratz *et al.* 2007] qui utilisent la maximisation du *Information Gain* comme stratégie de sélection de traits. Les auteurs ajoutent de nouveaux traits à ceux déjà proposés par [Dang & Palmer 2005], [Kohomban & Lee 2005]. Ces traits font à la fois appel à des informations contextuelles, morphosyntaxiques et sémantiques (entités nommées et hyperonymes). L'article ne donne pas le détail des meilleurs traits sélectionnés. Ce système obtient de meilleurs résultats que les systèmes évalués pour la tâche *English All Words - fine-grained* lors de Senseval-3 [Litkowski 2004b] et de SemEval mais les différences ne sont statistiquement significatives que par rapport à la *baseline*.

#### 2.4.1.3 Méthodes alternatives et systèmes d'apprentissage non supervisé

Néanmoins les approches non supervisées donnent également des résultats très intéressants. Nous rapportons ici les méthodes proposées par les travaux les plus remarquables de ce courant.

Certaines approches utilisent la correspondance des mots avec leur traduction dans une autre langue pour distinguer les différents sens d'un mot. C'est le cas des études de [Dagan *et al.* 1991] et [Dagan & Itai 1994] dans lesquelles la tâche de *sélection de mot-cible* en traduction automatique peut être assimilée à la tâche de désambiguïsation. L'idée est de traduire tous les mots d'une phrase par dictionnaire bilingue en conservant un certain nombre de candidats lorsqu'un mot source peut être traduit par différents mots-cibles. Pour chaque candidat cible $Cand_i$ et chaque mot ou candidat contextuel $Cont_j$ avec lequel il est en relation syntaxique $rel_r$, le système calcule alors les estimateurs de la probabilité que la relation syntaxique $rel_r(Cand_i, Cont_j)$ apparaissent dans la langue cible. Pour ne conserver que les candidats pour lesquels le taux de confiance est statistiquement significatif, Dagan borne le logarithme du rapport des chances (*odds ratio*) par la borne inférieure correspondant à l'estimateur moins l'intervalle de confiance pour un taux d'erreur $\alpha$ :

$$ln(\frac{p_1}{p_2}) \geq ln(\frac{\hat{p}_1}{\hat{p}_2}) - Z_{1-\alpha}\sqrt{var[ln\frac{\hat{p}_1}{\hat{p}_2}]}$$

avec $\hat{p}_1$ l'estimateur de probabilité du meilleur candidat et $\hat{p}_2$ celui du second meilleur candidat.

La sélection des meilleurs candidats s'effectue en choisissant séquentiellement les candidats qui maximisent le rapport des chances tout en dépassant le seuil fixé par l'intervalle de confiance. A chaque fois qu'un candidat est sélectionné, le système rejette les autres et les estimateurs de probabilité sont recalculés pour le choix du candidat suivant.

Dans ses travaux les plus connus, [Yarowsky 1995] expérimente différentes stratégies de désambiguïsation sur les deux assertions très fortes qu'il fait :

- *One sense per collocation* –les mots au voisinage d'un mot cible fournissent des indices très robustes sur le sens de celui-ci, moyennant la prise en compte de leurs distances relatives au mot cible, l'ordre et les relations syntaxiques les reliant ;

– *One sense per discourse* –le sens d'un mot cible est très régulier à l'intérieur d'un même document.

Les premiers résultats montrent que la simple assertion *One sense per discourse* peuvent permettre de donner une précision de quasiment 100% lorsqu'elle est applicable (en moyenne sur 50% des cas). Pour l'algorithme de désambiguïsation lui-même, Yarowsky propose dans un premier temps de regrouper toutes les occurrences d'un mot cible donné, de déterminer un mot discriminant de façon sûre chacune des classes de sens, et d'attribuer la classe correspondante à toutes les instances du mot-cible cooccurrant avec ce mot discriminant. Ensuite il entraîne sur ces données annotées un algorithme de listes probabilistes de décision et l'applique sur les données non annotées. Enfin, l'hypothèse de *One sense per discourse* vient infirmer certaines des attributions et surtout ajouter de nouvelles annotations aux instances de mot ambigu qui appartiennent au même document qu'un mot désambiguïsé. Cela permet de faire le lien avec des instances qui ne partageaient pas de contexte commun avec les mots déjà désambiguïsés (cf. le groupement *Cell* dans la figure 2.4). L'algorithme réitère ensuite l'apprentissage et l'extension faite par la stratégie *One sense per Discourse*. Les résultats donnés par cette étude montre même une amélioration par rapport au meilleur système supervisé du moment ([Schütze 1992]).

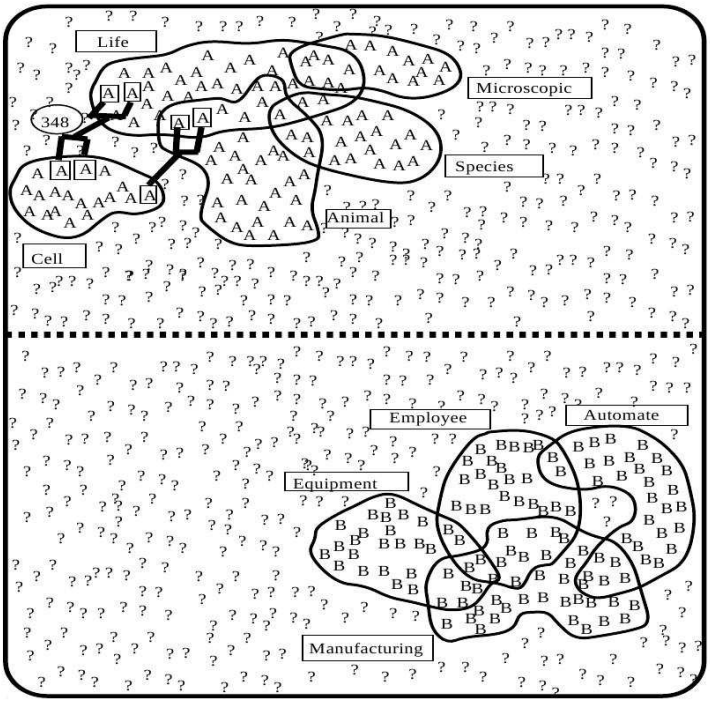



FIGURE 2.4 – One sense per discourse. Source : [Yarowsky 1995]

Il existe encore une quantité non négligeable de stratégies non supervisées proposées ces dernières années par tous les chercheurs s'étant intéressés à la désambiguïsation lexicale. Parmi ceux-ci on pourra citer [Dini *et al.* 2000] qui proposent un système à base de règles générées par les entrées du dictionnaire HECTOR augmentées d'informations issues de WordNet 1.5, [Yang & Powers 2006] qui utilisent la cohésion lexicale comme meilleur contexte de désambiguïsation et le thésaurus *Edimburgh Associative Thesaurus (EAT)*[25] pour étendre les contextes.

#### 2.4.1.4 Bootstrapping

Outre le système de [Yarowsky 1995], d'autres travaux se sont intéressés à des algorithmes itératifs pour la désambiguïsation, c'est le cas notamment des travaux de [Mihalcea & Moldovan 2001], [Mihalcea & Faruque 2004]. Leur stratégie est fondée sur l'observation suivante : dans l'histoire du WSD, un certain nombre d'algorithmes atteignent des précisions supérieures à 90% pour un nombre restreint de mots tandis que les évaluations menées sur l'ensemble du vocabulaire montrent que les système ne dépassent pas ces 90%. L'idée est donc de tenter de désambiguïser dans un premier temps, les mots pour lesquels la confiance en la désambiguïsation est forte et d'utiliser ensuite cette désambiguïsation comme une information complémentaire pouvant aider le traitement des autres mots pour lesquels on a une moins bonne précision.

Les auteurs définissent donc huit heuristiques d'attribution de sens, allant des plus sûres aux plus risquées, et appliquent celles-ci de façon itérative et séquentielle.

1. Entités nommées : les entités nommées détectées comme étant de type personne, lieu, et organisation sont remplacées par une étiquette indiquant ce type et le sens #1.
2. Mots monosémiques : les mots n'appartenant qu'à un seul synset dans WordNet sont étiquetés comme ayant le sens #1.
3. Indices contextuels : chaque mot-cible est associé avec le mot précédent et le mot suivant (à l'exclusion des déterminants et conjonction mais en conservant les prépositions). Si l'ensemble de leurs occurrences dans le corpus d'apprentissage ne correspond qu'à un seul sens #k et si le nombre d'occurrences du mot-cible avec l'étiquette sens #k dépasse un certain seuil, alors le mot-cible est étiqueté avec ce sens #k.
4. Contextes nominaux : les contextes nominaux de chaque sens sont pré-calculés sur SEMCOR dans une fenêtre de 10 mots pour tous les synsets hyperonymes d'un sens donné. Au moment de la désambiguïsation, le contexte du mot ambigu est comparé à ces ensembles de contextes et le sens ayant le plus de mots en commun est validé à condition que la différence entre le score du sens classé en premier et le score du sens suivant soit supérieure à un certain seuil.
5. Proximité sémantique de 0 dans WordNet avec un mot déjà désambiguïsé : les auteurs définissent une distance sémantique dans WordNet telle que deux mots appartenant au même synset

25. Le thésaurus EAT présente la particularité d'avoir été constitué en faisant appel à des sujets extérieurs. Il leur a été présenté des mots cibles pour lesquels ils devaient donner le mot auquel il pensait en premier.

sont à une distance de 0 tandis que les autres distances correspondent à la somme des nombres de chemins par la relation d'hyperonymie (et d'hyponymie). Si un des synsets du mot ambigu est à une distance 0 d'un mot déjà désambiguïsé dans l'ensemble du texte, alors ce synset est attribué au mot ambigu. A cette étape, seuls les synonymes d'un mot désambiguïsé sont donc annotés.

6. Proximité 0 avec un mot ambigu : la procédure est similaire sauf que cette fois-ci aucun des mots n'est préalablement désambiguïsé. L'heuristique est plus faible que la précédente et d'une complexité algorithmique plus grande. A cette étape, seuls deux synonymes peuvent être annotés.
7. Proximité 1 avec un mot désambiguïsé : cette étape prend en compte les mots liés par une relation d'hyperonymie/d'hyponymie avec un mot désambiguïsé.
8. Proximité 1 avec un mot ambigu : cette étape prend en compte les mots liés par une relation d'hyperonymie/d'hyponymie et non encore désambiguïsés.

Enfin pour certaines applications qui nécessitent un fort rappel (au détriment d'une perte de précision), les auteurs suggèrent d'utiliser la densité conceptuelle. La stratégie consiste à valider le sens qui partage le plus de voisins sémantiques avec les sens des mots précédent et suivant. Dans le cas où les catégories grammaticales sont différentes, l'heuristique passe par une étape supplémentaires dans la constitution des voisins sémantiques en utilisant les verbes figurant dans les définitions des sens du mot ambigu.

Les résultats montrent que 55% des mots sont désambiguïsés avec une précision de 92% (30% étant Entités nommées ou monosémiques) à l'aide des 8 premières heuristiques. En utilisant la densité conceptuelle, le système atteint un rappel de 90% et une précision de 82%.

#### 2.4.1.5 Désambiguïsation à partir de sens induits

Les méthodes de désambiguïsation à partir de sens induits sont encore assez rares puisque le domaine est assez nouveau. On trouve néanmoins plusieurs algorithmes dans la littérature, très fortement liés au type d'éléments clusterisés dans la phase d'induction.

On peut distinguer deux courants parmi ces approches non supervisées. Les approches de la première catégorie clusterisent certains des termes apparaissant dans le contexte des mots cibles sur un corpus d'apprentissage. Lors de la phase de désambiguïsation, un score est élaboré pour chaque cluster en fonction de l'appartenance ou non des termes du contexte du mot ambigu au cluster donné. C'est le cas de [Korkontzelos & Manandhar 2010] ou de [Véronis 2004] qui utilise directement l'arbre couvrant de poids minimal (*minimum spanning tree*) constitué dans la phase d'induction pour attribuer un score à chaque cluster de chaque mot ambigu rencontré. Si un mot figurant dans

le contexte du mot ambigu appartient à une des branches de l'arbre, alors il contribue au score du cluster auquel il est associé. Le cluster ayant le score le plus élevé est alors choisi comme sens du mot ambigu. Ces approches sont très similaires aux approches de type Lesk.

Les algorithmes de la deuxième catégorie définissent un modèle pour chaque instance du corpus d'apprentissage et clusterisent ces modèles d'instance. Le même processus est utilisé pour créer le modèle des instances à désambiguïser. Le cluster dont la distance avec le modèle de l'instance ambiguë est la plus faible est alors validé (il s'agit soit de la distance à partir du centroïde, soit de la moyenne des distances avec tous les éléments du cluster). C'est cette deuxième approche qui est adoptée par [Jurgens & Stevens 2010] (modèle de *Random Indexing*), par [Elshamy *et al.* 2010] (modèle de *Latent Dirichlet Allocation*), ou encore [Kern *et al.* 2010].

#### 2.4.1.6 Évaluation de la désambiguïsation utilisant des sens de mots induits

Les campagnes d'évaluation Semeval 2007 et 2010 se sont toutes deux intéressées à la tâche d'induction de sens en anglais. En 2010, [Manandhar *et al.* 2010] mettent en place un framework à trois étapes. Dans une première étape, les participants apprennent leurs sens de mots sur un corpus d'entraînement regroupant des instances de 50 verbes et noms cible en contexte. Dans un deuxième temps, ils utilisent leurs clusters de mots induits pour désambiguïser un corpus de test extrait de *Ontonotes* ([Hovy *et al.* 2006]). Les évaluateurs connaissent donc la vérité-terrain associée, annotée en sens *Ontonotes*.

C'est à la dernière étape que les évaluateurs procèdent à une évaluation automatique en deux parties. La première dite non supervisée, consiste à quantifier la qualité des clusters produits par le regroupement des instances annotées dans le corpus de test. Pour ce faire les évaluateurs utilisent la V-mesure [Rosenberg & Hirschberg 2007] mesurant à la fois l'homogénéité et la complétude des clusters, ainsi que la F-mesure paire à paire [Artiles *et al.* 2009], contrairement à leurs prédécesseurs [Agirre & Soroa 2007], qui utilisaient directement une F-mesure.

[Rosenberg & Hirschberg 2007] ont montré que la F-mesure n'est pas bien adaptée pour l'évaluation de la qualité d'un ensemble de clusters. En effet, d'une part elle ne capture pas l'homogénéité des parties de clusters contenant des éléments autres que la classe majoritaire dans un cluster donné. On voit dans les solutions $A$ et $B$ de la figure 2.5 que la F-mesure donne un score identique à ces deux ensembles de clusters. Pourtant, on voit par exemple que les clusters de la solution $B$ sont plus *purs* que ceux de la solution $A$ puisqu'ils contiennent des élément de seulement deux classes différentes contrairement aux clusters de A qui en contiennent trois. La V-mesure, en revanche, traduit bien une meilleure qualité de la solution $B$.

D'autre part, la F-mesure ne capture pas non plus la qualité des clusters secondaires contenant des éléments d'une classe déjà contenue majoritairement dans un autre cluster plus grand. On voit

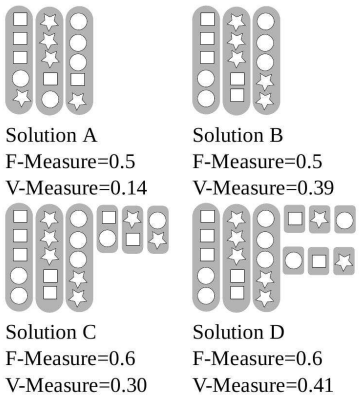



FIGURE 2.5 – Intérêt de la V-mesure comparé à la F-mesure. Source : [Rosenberg & Hirschberg 2007]

effectivement à la figure 2.5 que la F-mesure donne un score identique aux solutions $C$ et $D$ alors que la solution $D$ est parvenue à discriminer les différentes classes présentes dans les petits clusters, ce qui n'a pas été fait dans la solution $C$. On voit qu'à nouveau, la V-mesure rend bien compte de la meilleure qualité de $D$.

Concernant les résultats de cette évaluation, la meilleure V-mesure obtenue sur les noms est de 20,6 tandis que la meilleure V-mesure sur la désambiguïsation des verbes est de 15,6.

La deuxième partie de l'évaluation, dite supervisée, consiste à séparer le corpus de test en deux parties de 80% et 20%, la première partie servant à apprendre le mapping reliant les clusters utilisés par les candidats aux sens *Ontonotes*. La séparation est reproduite aléatoirement cinq fois, la mesure de rappel communément utilisée dans les tâches de WSD peut alors être calculée sur chacun des échantillon. La mesure finale est alors constituées par la moyenne des cinq mesures précédemment calculées.

Le système obtenant le meilleur rappel sur les noms est le même que celui ayant obtenu la meilleure V-mesure sur les noms. Il obtient un rappel de 59,4% sur les noms. En revanche, le système donnant le meilleur rappel sur les verbes obtient un classement beaucoup plus modeste. Le meilleur rappel sur les verbes est de 69,1%.

#### 2.4.1.7 Bilan

Comme le rappellent [Stevenson & Wilks 2001], les systèmes peuvent bénéficier d'un grand nombre de connaissances externes aux données du problème. Les sources de connaissances utilisées en désambiguïsation lexicale sont très variées.

En effet, la simple information de catégorie grammaticale suffit parfois à désambiguïser le sens des mots comme dans l'exemple *well[adj] / well[subst]* en anglais ou le premier mot réfère à l'adjectif français *bien* tandis que le deuxième réfère au *puit*. Nous avons le même cas en français pour *vive[adj]*

/ *vive[subst]* où le premier mot réfère à l'adjectif *vif* au féminin tandis que le second réfère au *poisson* de ce nom.

On peut ensuite faire appel à des connaissances plus sémantiques de restrictions de préférence. Dans le cas de *The bat sleeps.*, si le système sait que le sujet de *sleep* doit être animé, alors il saura qu'il s'agit de la chauve-souris et non de la *batte de baseball* ou de la *raquette de ping-pong*.

Enfin, les systèmes peuvent également faire appel à des connaissance de type pragmatique. Pour traiter la phrase *He buys a bat in a sport shop.*, le système peut avoir accès à une base de connaissance externe l'informant que *bat* et *sport shop* sont fortement liés lorsque *bat* ne fait pas référence à la chauve-souris.

En ce qui concerne les frameworks d'évaluation et de comparaison des différents systèmes, [Gale *et al.* 1992b] proposent une évaluation par pseudo-mots qui présente l'avantage de ne pas nécessiter de corpus annoté. L'idée est de prendre des paires de mots monosémiques du vocabulaire et de simuler une ambiguïté sur chacune de leurs instances respectives en remplaçant celles-ci par le pseudo-mot constitué des deux mots source. L'objectif est alors pour le désambiguïsateur de déterminer lequel des deux mots était présent à l'origine. Nous ne retiendrons pas cette méthodologie car nous partageons l'opinion de [Gaustad & Groningen 2001] : la performance varie trop en fonction des pseudo-mots choisis et la tâche n'est pas la même que la désambiguïsation de mots réels. Les différentes méthodes de choix de pseudo-mots permettant d'être exhaustif et objectif sur la tâche évaluée (par exemple [Nakov & Hearst 2003]) ne sont pas suffisamment satisfaisantes ni facilement applicables à nos données (y compris pour l'évaluation de l'induction de sens seule).

Plus récemment, la première campagne d'évaluation Senseval eut lieu en 1998 ([Kilgarriff & Rosenzweig 2000], [Segond 2000]) et fut suivie de 2 autres campagnes du même nom. Ces campagnes menèrent à une description standard de la tâche ainsi qu'à la mise au point d'une méthodologie d'évaluation reconnue.

Enfin, une des grandes questions soulevées par l'ensemble de ces travaux reste la granularité des référentiels de sens utilisés. Dans les campagnes Senseval, les sens utilisés ont souvent été ceux du dictionnaire HECTOR puis ceux de WordNet (versions successives). Les accords inter-annotateurs rapportés pour l'annotation manuelle sont relativement faibles et bornent la performance des systèmes. Dans le cadre de Senseval-1 sur le français (dictionnaire petit Larousse), [Segond 2000] rapporte un taux d'accord (calculé deux à deux et pondéré par le coefficient de Dice) de 69% dont 63% pour les verbes, tandis que sur la tâche *English All words* de Senseval-3 (synsets WordNet 1.7), [Snyder & Palmer 2004] rapporte un accord de 72.5% en moyenne dont 67.8% pour les verbes.

A priori la distinction de sens est souvent trop ambiguë pour l'humain, soit le mot emploie deux des sens du référentiel à la fois, soit un sens qui serait entre les deux. Dans tous les cas, beaucoup

prônent l'utilisation d'une distinction de sens à plus grosse granularité que la granularité la plus fine de WordNet. En particulier dans un cadre plus applicatif, il y a peu de chances qu'une granularité si fine soit intéressante en soi (d'autant plus que la précision des systèmes est moins bonne).

### 2.4.2 L'annotation en rôles sémantiques

Comme nous l'avons vu à la section 2.2.1.3, l'annotation en rôles sémantiques consiste à annoter les prédicats d'une phrase ainsi que leurs arguments, avec des étiquettes sémantiques d'une ressource de référence. Cette tâche peut être divisée en quatre sous-tâches même si celles-ci sont fortement corrélées. Il s'agit tout d'abord de détecter les prédicats dans une phrase, puis d'identifier la frame déclenchée par rapport à une ressource externe de référence. Dans un deuxième, il faut alors détecter les arguments du prédicat, et également les identifier par rapport aux étiquettes fournies dans la ressource de référence.

Considérons que l'on souhaite annoter un texte avec la ressource de référence FrameNet, la phrase exemple présentée dans l'introduction serait alors annotée de la façon suivante :

| | | | |
|---|---|---|---|
| | **Activity_finish** | [Time] | |
| [Activity] | | | [Result] |
| La table ronde «radio-fréquences, santé, environnement» | **s'est achevée** | ce lundi | avec une dizaine de pistes mais sans aucune mesure forte. |

#### 2.4.2.1 Corpora

Comme nous le verrons dans la suite de cette section, les méthodes les plus performantes en annotation de rôles sémantiques sont celles qui utilisent un apprentissage supervisé. C'est pourquoi nous donnons un bref aperçu des différents corpus annotés existants.

Le premier corpus est celui de FrameNet (déjà décrit dans 2.2.1.3). Il est constitué de 150 000 phrases du *British National Corpus (BNC)* annotées manuellement à l'aide des *frames* répertoriées dans la base FrameNet et de leurs rôles associés, les rôles étant spécifiques aux frames.

PropBank ([Palmer *et al.* 2005]) est un corpus extrait du Wall Street Journal (nouvelles financières en anglais américain) où les verbes sont annotés avec leurs arguments sémantico-syntaxiques.

Enfin, il existe un corpus de langue allemande annoté manuellement. Ce corpus, SALSA [26], est décrit dans [Burchardt *et al.* 2006]. Il contient actuellement 20 000 instances verbales annotées par les frames du FrameNet de Berkeley ainsi qu'un sous-ensemble de frames spécifiques ajoutées au FrameNet allemand.

Il existe aussi des méthodes de projection d'annotations d'une langue à l'autre (cf. 2.3.1.2). Des corpus en français, allemand, italien ont ainsi été obtenus, mais leur taille reste assez faible et leur

26. http://www.coli.uni-saarland.de/projects/salsa/corpus/

précision également.

#### 2.4.2.2 Métriques

Dans le cadre des campagnes d'évaluation, Senseval-3 a également proposé une tâche d'annotation en rôles sémantiques. [Litkowski 2004a] rapporte le protocole et les différentes conclusions de cette session d'évaluation. Les objectifs, la baseline et les mesures utilisées pour l'évaluation furent fournis par l'étude antérieure de [Gildea & Jurafsky 2002]. Les corpus de vérité-terrain sont issus du corpus FrameNet et annotés manuellement comme décrit dans [Johnson *et al.* 2003]. Les mesures d'évaluation sont la précision, le rappel et le recouvrement des *frames* identifiées par le système candidat avec celles annotées manuellement. La tâche consiste à identifier les rôles (*Frames Elements* ou *FE*) sur les données de test et à les étiqueter (les frames et prédicats déclencheurs des frames sont donnés).

Chaque *frame* pour laquelle un système a tenté d'annoter les différents rôles est comparée à l'annotation humaine. Les réponses correspondantes sont selon le cas rejetées ou validées comme bonnes réponses avec un score de recouvrement. Les mesures utilisées sont les suivantes :

- Précision = $\frac{Nombre\ de\ bonnes\ réponses}{Nombre\ de\ réponses\ tentées}$
- Rappel = $\frac{Nombre\ de\ bonnes\ réponses}{Nombre\ de\ FE\ du\ test\ set}$
- Recouvrement = $\frac{1}{|\{FEs\}|} \cdot \sum_{fe \in \{FEs\}} rec(fe)$
  avec $rec(fe) = \frac{Nombre\ de\ lettres\ communes\ couvertes\ par\ les\ annotations\ automatique\ et\ manuelle}{Nombre\ de\ lettres\ dans\ annotation\ du\ test\ set}$
- Tentatives = $\frac{Nombre\ de\ réponses\ FE\ générées}{Nombre\ de\ FE\ du\ test\ set}$

Les résultats rapportés par [Litkowski 2004a] sont présentés dans le Tableau 2.3. Le meilleur système dans les deux tâches proposées est celui de [Bejan *et al.* 2004] utilisant un classifieur supervisé de type SVM.

| **Tâche** | **Système** | **Précision** | **Recouvrement** | **Rappel** | **Tentatives** |
|---|---|---|---|---|---|
| Avec bordures de *FE* | Meilleur système | 94,6% | 94,6% | 90,7% | 95,8% |
| | Moyenne des 20 systèmes candidats | 80,3% | | 75,7% | |
| Sans bordures de *FE* | Meilleur système | 89,9% | 88,2% | 77,2% | 85,9% |
| | Moyenne des 20 systèmes candidats | 59,5% | | 48,1% | |

TABLE 2.3 – Résultats de l'évaluation Senseval-3.

La campagne SemEval menée en 2007 a proposé deux tâches liées au SRL. La première est associée à la tâche de désambiguïsation lexicale décrite par [Pradhan *et al.* 2007] et consiste à annoter les

différents arguments PropBank pour un certain nombre de lemmes. La deuxième décrite par [Baker *et al.* 2007] consiste à extraire les prédicats évoquant des frames de FrameNet ainsi que les différentes syntagmes réalisant les rôles associés.

Le meilleur système pour la tâche d'annotation FrameNet est celui de [Johansson & Nugues 2007a] (pour seulement deux participants à la tâche complète !). Le système décrit entraîne des SVM sur des relations de dépendances syntaxiques, ce qui est assez novateur pour un système de SRL. De plus, le système utilise un dictionnaire enrichi décrit dans [Johansson & Nugues 2007b] pour l'appartenance des différentes unités lexicales aux frames. Les précisions de ce système pour les différents corpus varient entre 54 et 67 % pour des rappels respectifs de 36 et 46 %. Les résultats de cette évaluation sont moins élevés que ceux de la tâche précédente. Cela s'explique par la plus grande variété des corpus ainsi que l'extension de la tâche aux unités lexicales non présentes dans FrameNet.

Nous ne rapportons pas ici en détails les résultats des évaluation CONLL ([Surdeanu *et al.* 2008], [Hajič *et al.* 2009]) car elles concernent plus particulièrement l'annotation en arguments PropBank qui n'est pas le référentiel retenu pour notre système. Nous pouvons cependant prendre note du fait que la tâche de *joint learning* suggérée pour une meilleure prise en compte de l'interface syntactico-sémantique ne semble pas encore avoir fait ses preuves. En effet, les systèmes cherchant à résoudre la tâche d'annotation en dépendances syntaxiques en même temps que la tâche d'annotation en dépendances sémantiques ne donnent pas globalement de meilleurs résultats que les systèmes participant à la tâche de SRL pure (bien qu'il existe des cas particuliers pour certaines langues mais c'est un phénomène ponctuel et non significatif). Notons aussi que 3 des 4 meilleurs systèmes de la campagne 2009 utilisent des modèles à maximum d'entropie (ME).

La campagne SemEval-2 menée en 2010 n'a proposé qu'une tâche de SRL prenant en compte les rôles répartis sur plusieurs phrases (*Linking events and their Participants in Discourse*), tâche sur laquelle ne nous attarderons pas car elle dépasse le cadre de notre étude.

#### 2.4.2.3 Des systèmes supervisés pour la langue anglaise

La très grande majorité des systèmes d'annotation automatique de rôles sont des systèmes d'apprentissage supervisés. On peut remarquer notamment l'article fondateur de [Gildea & Jurafsky 2002].

L'apprentissage se fait sur le corpus issu du projet FrameNet contenant 50 000 phrases du BNC annotées en rôles sémantiques. Les différentes caractéristiques utilisées pour l'apprentissage sont les types de syntagme (NP, PP, ADVP, PRT, SBAR, S), la catégorie parente des NP (S ou VP), le chemin du *parse tree*, la position du constituant (avant ou après le prédicat), la voix (active ou passive) et le lemme tête du constituant. Pour chacune de ces caractéristiques et quelques combinaisons de celles-ci, le système calcule sur les corpus d'apprentissage les probabilités conditionnelles qu'un constituant remplisse un rôle donné. Chacune des caractéristiques a des propriétés de couverture et de précision différentes. L'estimation des probabilités conditionnelles prenant en compte l'ensemble

des caractéristiques est effectuée dans un premier temps par une combinaison linéaire des probabilités unitaires. Deux stratégies sont utilisées pour le choix des poids d'interpolation. La première méthode consiste à attribuer des poids égaux lorsque la condition est présente dans le corpus d'apprentissage. La seconde stratégie utilise l'algorithme de *Expectation Maximization (EM)* ([Jelinek & Mercer 1980]) pour optimiser ces poids. Les auteurs évaluent aussi les performances d'une estimation de probabilité en considérant la moyenne géométrique des probabilités unitaires, et deux estimations utilisant les probabilités unitaires les plus spécifiques avec un *back-off* vers une probabilité unitaire plus générale dans les cas où il n'y a pas de données d'apprentissage pour une condition donnée. Les meilleurs résultats sur le corpus d'apprentissage sont obtenus avec l'algorithme *back-off* en combinaison linéaire à poids égaux tandis que les meilleurs résultats sur le corpus de test sont obtenus avec la simple combinaison linéaire pondérée par l'algorithme EM.

L'article aborde également la tâche de détection des *frame elements*. Les mêmes caractéristiques sont utilisées pour le calcul de la probabilité qu'un constituant soit effectivement un *frame element*.

L'étude de modules complémentaires pour l'annotation de rôles sémantiques est également décrite. Le clustering pratiqué sur les noms du corpus d'apprentissage permet de calculer des probabilités pour les têtes de constituants n'apparaissant pas dans les données d'apprentissage. Les probabilités conditionnelles liées au clustering seul couvrent presque l'intégralité de l'ensemble de test et sont légèrement moins précises que précédemment. En revanche, une fois combinées à l'ensemble des autres caractéristiques, la précision s'améliore également. Une autre stratégie utilisant WordNet s'avère également efficace tout en restant moins bonne que celle utilisant le clustering. Les auteurs étudient aussi la prise en compte de contraintes de *Frame Elements Group*, qui semble être nécessaire mais redondante avec la prise en compte des cadres de sous-catégorisation.

Le système dans son intégralité atteint une précision de 82% quand les syntagmes et propositions qui remplissent les différents rôles sont pré-segmentés et une précision de 65% pour un rappel de 61% lorsque le système doit lui-même segmenter les différents constituants. Ces travaux adressent également la généralisation nécessaire lorsque le système rencontre un prédicat n'existant pas dans les données d'entraînement.

Enfin, l'article traite également de la généralisation aux prédicats non connus, aux frames non connues et aux domaines sémantiques (de FrameNet) non connus. Les évaluations sont menées en enlevant un certain nombre de prédicats aux données d'apprentissage pour les mettre dans les données de test. De la même façon, pour les frames non connues, on enlève une frame aux données d'apprentissage et le système utilise toutes les autres pour apprendre, ceci étant effectué pour chacune des frames du corpus. L'interpolation proposée pour l'estimation de la probabilité reste robuste pour la généralisation cross-prédicats mais les performances sont nettement dégradées pour les généralisations cross-frames et cross-domaines.

Depuis ces travaux fondateurs, beaucoup de systèmes partagent la structure de cet algorithme en partitionnant la tâche en sous-tâches telles que : *Détection/Identification du prédicat cible*, *Désambiguïsation de la frame déclenchée*, *Détection/Identification des arguments remplissant les rôles* et *Classification de ces arguments*. La très grande majorité de ces systèmes utilisent des classifieurs supervisés exploitant un ensemble de caractéristiques comprenant celles suggérées par [Gildea & Jurafsky 2002]. Certains travaux reviennent néanmoins sur les différentes caractéristiques utilisées en montrant qu'elles pourraient être plus informatives en fonction de la sous-tâche traitée. C'est le cas notamment de [Xue & Palmer 2004] ainsi que [Pradhan *et al.* 2008]. Ces derniers identifient notamment le chemin syntaxique du prédicat à l'argument ainsi que le chemin partiel (du plus petit noeud commun à l'argument) comme les traits les plus influents sur la tâche de détection, le prédicat ne fournissant que très peu d'information influente. En ce qui concerne la classification des arguments, les traits les plus saillants semblent être des éléments moins structurels et plus lexico-sémantiques, telles que la tête, le premier mot et le dernier mot de l'argument considéré ainsi que le prédicat lui-même.

D'après l'état de l'art rédigé par [Das *et al.* 2010], les principaux systèmes développés depuis [Gildea & Jurafsky 2002] sont ceux de : [Fleischman *et al.* 2003] qui utilisent pour la première fois des modèles à maximum d'entropie dans le but de détecter et de classifier les arguments pour une frame donnée, [Toutanova *et al.* 2005] qui entraînent un modèle log-linéaire incorporant des traits globaux contraignant les labels entre eux, [Thompson *et al.* 2004] qui proposent un modèle génératif, [Shi & Mihalcea 2004] qui développent un système à base de règles et [Erk & Padó 2006] qui étudient l'intérêt d'un classifieur Bayes naïf. D'autres emploient des classifieurs à base de SVM (*Support Vector Machine*) et obtiennent souvent de très bons résultats, c'est le cas entre autres de [Bejan *et al.* 2004], [Bejan & Hathaway 2007], [Moldovan *et al.* 2004], [Giuglea & Moschitti 2006], [Johansson & Nugues 2007a], [Pradhan *et al.* 2008].

[Shen & Lapata 2007] proposent une approche par graphes dans laquelle la sortie de l'analyseur n'est pas un rôle unique par *frame element* mais un ensemble de rôles pondérés. La tâche de SRL est modélisé comme étant un problème de *minimum weight bipartite edge cover*, où la première partie du graphe bipartite correspond aux différents *frame elements*, la seconde partie correspond aux rôles sémantiques et le score associé aux arêtes est donné par une similarité de sous-séquences entre chemins syntaxiques. La tâche de SRL elle-même n'est pas comparée aux évaluations classiques de type SemEval du fait de sa sortie multiple. En revanche, la sortie du système est utilisée dans une tâche de Questions/Réponses dont les résultats sont évalués sur les données TREC QA 2002 à 2005 et comparés à deux approches baseline montrant ainsi une amélioration significative (cf. 2.5.3.2).

Enfin, [Das *et al.* 2010] proposent un modèle probabiliste obtenant de meilleurs résultats que [Johansson & Nugues 2007a] (le meilleur système de SemEval 2007), pour chacune des sous-tâches identifiées. Les principaux apport de ce travail sont au nombre de trois : l'utilisation d'un modèle à variable latente permettant la désambiguïsation de mots ne figurant pas dans le lexique de FrameNet,

un modèle unifié pour l'identification et la classification des arguments, et enfin une contrainte empêchant les arguments d'un même prédicat de se chevaucher.

#### 2.4.2.4 Moins de supervision, plus de langues

Les méthodes présentées jusqu'ici nécessitent toutes un très grand nombre de données d'apprentissage. Or il n'existe que deux corpus annotés pour l'anglais : un extrait du Wall Street Journal pour PropBank pour lequel le domaine est très typé *Nouvelles financières* et un extrait du British National Corpus pour FrameNet, corpus plus généraliste mais dont les prédicats annotés appartiennent à un ensemble prédéfini de frames. Pour d'autres langues que l'anglais, de tels corpora sont encore plus rares.

Nous proposons donc de nous concentrer sur l'étude des méthodes faiblement supervisées afin de nous en inspirer pour pouvoir traiter la langue française.

Les premières tentatives d'annotation automatique en rôles sémantiques de façon non supervisée sont assez récentes. On ne trouve que quelques articles dans la littérature. Voici les différentes méthodes que nous avons pu trouver.

La première méthode non supervisée présentée par [Swier & Stevenson 2004] utilise un modèle probabiliste et VerbNet comme lexique de prédicats. Pour l'identification du prédicat et des rôles à annoter, la méthode extrait une liste de candidats constitués d'un verbe associé à un schéma de rôles syntaxiques (cet association définit une *frame*). Cette liste est donnée directement par l'entrée du verbe de la phrase dans VerbNet. Un score de correspondance est ensuite calculé entre les schémas de rôles syntaxiques des *frames* et l'analyse syntaxique de la phrase à annoter. La *frame* donnant le meilleur score est conservée pour l'annotation de la phrase donnée, et les rôles sont annotés suivant la correspondance syntaxique.

[Swier & Stevenson 2004] définissent ensuite trois niveaux de modèles probabilistes mis à jour itérativement. Ces modèles probabilistes sont fondés sur le modèle le plus spécifique $P(r|v, s, n)$ de la probabilité d'attribution d'un rôle en fonction du verbe utilisé, de la fonction syntaxique et du nom utilisé (tête du syntagme à annoter). Les deux autres niveaux se rendent indépendants d'une ou plusieurs conditions tout en généralisant une des autres conditions (classes d'équivalence définies pour les verbes, les fonctions syntaxiques, et les têtes de syntagme). L'attribution des rôles est ensuite effectuée par itération si les probabilités des candidats sont au-dessus d'un certain seuil $\theta 1$ et que le *log ratio* du meilleur candidat comparé au second est au-dessus d'un certain seuil $\theta 2$. On réitère jusqu'à ce qu'il n'y ait plus de nouvelles assignations. Puis $\theta 2$ est décrémenté et on recommence, jusqu'à ce qu'il soit nul. Les cas d'égalité sont ensuite résolu par le dernier modèle probabiliste.

Ils atteignent un taux de réduction d'erreur de 50-65 % par rapport à leur baseline consistant à utiliser la probabilité conditionnelle $P(r|sc)$, probabilité d'attribution du rôle en fonction de la classe de la fonction syntaxique. Ceci correspond à une précision d'annotation des rôles de 87 %.

Ces résultats sont très satisfaisants mais il faut retenir que la méthode exploite très fortement le fait que les rôles définis par VerbNet sont les mêmes quel que soit le verbe déclencheur et il n'existe que très peu de systèmes utilisant les classes de VerbNet pour cette tâche.

Un des premiers travaux à soulever l'intérêt de l'apprentissage semi-supervisé, [Thompson 2004] applique l'algorithme fondé sur l'algorithme *EM* (espérance-maximisation) de [Dempster *et al.* 1977] mais n'obtient pas de résultats significatifs.

[He & Gildea 2006] comparent deux méthodes semi-supervisées, le *self-training* utilisant un classifieur de type *Maximum Entropy* et le *co-training* sur des classifieurs de type *listes de décision* afin d'annoter des phrases contenant également des *frames* inconnues. Les rôles de FrameNet sont généralisés à un ensemble de 15 rôles thématiques afin que l'apprentissage puisse être effectif. La plus performante des deux méthodes (*self-training*) n'apporte pas de gain significatif lors de l'introduction des données inconnues dans les données d'apprentissage, et le taux de précision pour les données inconnues reste stationnaire autour de 35 %.

[Fürstenau & Lapata 2009b] proposent une méthode semi-supervisée pour l'annotation des rôles (sans traiter de l'identification des *frames*) dans laquelle un sous-ensemble d'exemples de phrases est annoté pour un sous-ensemble de frames prédéfinies. La méthode permet d'augmenter la taille du corpus d'apprentissage automatiquement pour les frames données et cela a pour conséquence l'augmentation de la F-mesure comparé à une méthode supervisée n'utilisant que le sous-ensemble initial comme données d'apprentissage. L'article décrit en outre l'obtention des différents paramètres de l'algorithme et leur influence respective. Cette approche fonctionne bien pour les frames données mais elle nécessite encore l'existence d'un corpus initial où toutes les *frames* et rôles que l'on souhaite annoter automatiquement figurent et y soient manuellement annotés un certain nombre de fois.

Dans un travail complémentaire, [Fürstenau & Lapata 2009a] étendent leur étude au traitement des verbes ne figurant pas dans la liste des unités lexicales déclencheuses de *frames* dans FrameNet en utilisant une similarité calculée à l'exécution avec les *frames* existantes, comme proposé par [Pennacchiotti *et al.* 2008] pour l'enrichissement de ressources. Les mesures de similarité sont revues pour une meilleure performance et les auteurs rapportent des mesures améliorées pour les *frames* de moyennes et basses fréquences. Les *frames* nécessitent toujours d'être connues et manuellement annotées dans le corpus d'initialisation.

Les travaux de [Deschacht & Moens 2009] proposent un nouveau modèle pour l'apprentissage de similarités sur des textes non annotés et l'appliquent à la tâche de SRL semi-supervisée. Le *Latent Word Language Model (LWLM)* est un nouveau modèle d'espace sémantique qui se fonde sur les contextes de latence des mots apparaissant dans une fenêtre de taille 2 autour des mots à caractériser. Cette mesure de similarité est utilisée dans des algorithmes semi-supervisés en tant

que trait supplémentaire à un algorithme supervisé et comme mesure de similarité sémantique pour un algorithme similaires à celui de [Fürstenau & Lapata 2009b]. L'évaluation faite sur le corpus de PropBank montre que le *LWLM* en tant que trait donne de meilleurs résultats que ceux de [Goodman 2001] quelle que soit la taille du corpus initial. En tant que similarité sémantique dans un algorithme de type [Fürstenau & Lapata 2009b], plus la taille du corpus initial est grande, plus les performances sont améliorées. Cependant, les résultats restent inférieurs à ceux d'une méthode entièrement supervisée lorsque la taille du corpus dépasse 20 % de la taille du corpus PropBank. Ces travaux ne présentent pas en soit de nouvel algorithme semi-supervisé mais apportent une mesure de similarité intéressante.

Certains s'intéressent aussi à la résolution non supervisée de sous-tâches de SRL comme [Grenager & Manning 2006] qui traitent le problème de la classification des arguments par rapport aux rôles de PropBank, ou [Abend *et al.* 2009] qui présentent une méthode non supervisée pour l'identification préalable des arguments. Deux heuristiques sont définies : la première se fonde sur le fait que dans PropBank, 86 % des arguments figurent dans les clauses minimales du prédicat. Les auteurs décrivent donc un ensemble de règles visant à déterminer les clauses minimales dans lesquelles figurent la majorité des arguments.

La deuxième part du principe que l'argument d'un prédicat n'apparaît pas pour la première fois dans son contexte. Les auteurs pré-calculent donc sur un corpus externe le PMI (*Pointwise Mutual Information*) des prédicats (lemme+PoS) *versus* les têtes de mots apparaissant dans les clauses minimales. Si les fréquences d'apparitions des prédicats et arguments candidats sont supérieurs à un certain seuil dans le corpus externe, alors le PMI est utilisé après la première heuristique comme filtre pour l'identification des arguments.

Les paramètres fournissant la meilleure précision permettent d'atteindre une précision de 56 % (avec un rappel à 40 %) contre une précision de 47% pour la baseline. Ces résultats sont intéressants, mais ils restent encore bien en deçà de ceux des meilleurs systèmes supervisés qui atteignent une précision de 81%. Dans la suite logique de ces travaux, [Abend & Rappoport 2010] proposent également une méthode non supervisée de classification des arguments en *core* ou *non-core*, qui est une première étape vers la classification complète des arguments.

#### 2.4.2.5 Bilan

La tâche d'annotation en rôles sémantiques s'est beaucoup développée avec l'apparition de la ressource FrameNet à partir de 1998. Les premières méthodes d'annotation que l'on trouve dans la littérature sont principalement des méthodes supervisées. Ainsi, pour traiter d'autres langues que l'anglais pour lesquelles il n'existe pas de corpus d'apprentissage annoté à la main, les alternatives sont assez restreintes. Certains proposent des projections d'annotations des corpus anglophones à des corpus d'autres langues afin de disposer de corpus d'apprentissage (même imparfait) pour les autres

langues ([Padó & Lapata 2005]), d'autres annotent manuellement de très petits corpus et tentent d'apprendre à partir de ceux-ci par des méthodes incrémentales [Fürstenau & Lapata 2009b]. Il n'existe à l'heure actuelle que très peu de systèmes complètement non-supervisés pour la tâche d'annotation sémantique de rôles.

### 2.4.3 Conclusions

Les méthodes les plus performantes pour les deux tâches d'analyse sémantique identifiées sont celles utilisant un apprentissage supervisé. Ces approches sont applicables pour les langues pour lesquelles on dispose de corpus annotés, ce qui est généralement le cas pour l'anglais. Peu de corpus annotés sont disponibles pour le français. Ainsi, nous nous tournerons vers des méthodes peu ou non supervisées pour la désambiguïsation et l'annotation en rôles sémantiques, avec pour objectif de proposer de nouvelles approches dépassant les performances des méthodes non supervisées connues à ce jour.

Nous verrons dans la section suivante à quelles problématiques applicatives ces deux tâches d'analyse sémantique peuvent être utiles.

## 2.5 La sémantique en recherche d'information

Les résultats des recherches en sémantique lexicale pour les tâches que nous venons de voir sont peu à peu intégrés dans des applications de recherche d'information. Les principaux intérêts que nous voyons dans une telle démarche portent sur trois points majeurs que sont l'indexation par sens, l'extension ou le raffinement des requêtes utilisateurs et l'usage des rôles sémantiques dans les systèmes complexes de questions-réponses. Nous étudions donc plus en détails l'intégration de modules de traitements sémantiques sur ces trois thématiques.

### 2.5.1 Indexation par sens

Une des premières expériences significatives d'indexation de documents non plus par mots mais par sens est l'étude de [Voorhees 1993] dans laquelle 5 collections de documents sont indexés par le sens de leur mots désambiguïsés selon les synsets de WordNet. Les requêtes envoyées au système sont à leur tour désambiguïsées. L'expérience montre qu'un petit nombre de requêtes sont améliorées par cette désambiguïsation. En revanche, dans la majorité des cas, le peu de contexte des requêtes courtes conduit à une désambiguïsation erronée des mots polysémiques ou à une non-décidabilité de désambiguïsation, les deux cas menant à une performance réduite de la correspondance entre requête et documents pertinents et par suite à une précision plus faible. [Voorhees 1993] rejoint cependant l'avis de [Krovetz & Croft 1992] en affirmant qu'une désambiguïsation correcte contribuerait à l'amélioration de la pertinence des résultats.

[Sanderson 2000] passe en revue les différentes expériences dans le domaine et conclue par les trois points suivants. L'ambiguïté des mots n'apparaît pas comme un problème majeur en recherche d'information dû au fait que certains sens de mots sont prévalents sur les autres (*skewed distribution* = un sens apparaît dans plus de 80% des cas), et que les collocation des autres mots de la requête (s'ils existent et que les documents sont suffisamment grands) contrebalancent les potentiels problèmes d'ambiguïté) ([Krovetz & Croft 1992]). La précision d'un désambiguïsateur doit être très élevée (supérieure à 90%) pour qu'elle apporte un bénéfice à un système de recherche d'information. Une représentation du sens des mots par rapport à un simple dictionnaire semble être d'une granularité trop fine pour des applications de recherche d'information.

D'autres recherches mènent également à des conclusions mitigées sur l'intérêt de la désambiguïsation lexicale (WSD) en recherche d'information, mêmes si les arguments ne sont pas toujours les mêmes. Dans les conclusions de son chapitre de thèse sur l'évaluation des synonymes et de la désambiguïsation dans les systèmes de recherche documentaire, [Loupy (de) 2000] souligne le fait que le système de WSD utilisé est d'une précision inférieure à 90% mais ne dégrade pas pour autant les résultats de la recherche documentaire. Mais *in fine* il ne parvient pas non plus à affirmer l'intérêt de la désambiguïsation en recherche d'information :

> Même si l'apport des synonymes et de la désambiguïsation ont été montrés dans les expériences présentées ici, les améliorations sont faibles et parfois même non significatives.

Cette opinion n'est pourtant pas partagée par tous, [Véronis 2004] conclut ses travaux sur la découverte de sens par l'algorithme HyperLex, la désambiguïsation lexicale, la visualisation et la navigation dans les résultats, en suggérant que la désambiguïsation ne peut pas être néfaste à la recherche d'information :

> Enhancements are of course possible, but this study already seems to cast doubt on the idea that disambiguation techniques are useless if not detrimental in IR.[27]

Il préconise néanmoins de procéder plutôt à une suggestion de sens détectés dans le corpus plutôt que d'essayer de désambiguïser une requête qui n'a qu'un contexte très pauvre, voire inexistant.

Plus récemment, la tâche *Robust WSD Track* ([Agirre *et al.* 2009]) de la campagne CLEF 2009 proposait un corpus de documents et de requêtes dont les mots avaient été préalablement désambiguïsés en synsets WordNet par deux systèmes de désambiguïsation état de l'art. Chaque participant était tenu de fournir au minimum un système n'exploitant pas les données de désambiguïsation et un système les exploitant, de sorte que l'on puisse quantifier l'intérêt d'une telle information. Les systèmes ayant participé à cette évaluation ont usé de deux stratégies différentes et complémentaires. Les uns ont tenté une indexation par sens de façon similaire aux systèmes décrits dans cette section, les autres

27. Des améliorations sont bien sûr possibles, mais cette étude semble déjà jeter le doute sur l'idée que les techniques de désambiguïsation sont inutiles si ce n'est nuisibles à la Recherche d'Information.

ont proposé des systèmes d'extension de requêtes. Nous introduisons donc maintenant les différentes stratégies existant dans la littérature concernant cette sous-tâche de la recherche d'information.

### 2.5.2 La reformulation de requête...

Si l'on tente de distinguer plusieurs générations dans l'histoire des moteurs de recherche de documents, on peut tracer à gros traits quatre générations. Les moteurs de première génération se contentent de renvoyer les résultats correspondant à la requête booléenne de l'utilisateur sans proposer de finesse dans leur classement.

Dans les années 60, [Salton *et al.* 1975] proposent une approche vectorielle dans laquelle chaque document et chaque equête sont représentés par un vecteur. Une fois cette modélisation disponible, une recherche de document est alors effectuée en trouvant les vecteurs de documents les plus proches du vecteur requête par la similarité cosinus.

Chaque coordonnée d'un vecteur correspond à un mot du vocabulaire et la valeur qui lui est attribuée dépend de la fréquence du terme dans le document et de l'inverse de la proportion du nombre de documents contenant au moins une fois le terme donné.

Cette pondération est appelée *TfIdf* (*Term Frequency - Inverse Document Frequency*).

Soit le document $d_j$ et le terme $t_i$, la fréquence du terme (*Term Frequency*) dans le document est donnée par :

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

avec $n_{i,j}$ est le nombre d'occurrences du terme $t_i$ dans $d_j$.

La fréquence inverse de document (*Inverse Document Frequency*) est une mesure de l'importance du terme dans l'ensemble du corpus. Elle consiste à calculer le logarithme de l'inverse de la proportion de documents du corpus qui contiennent le terme, soit :

$$\mathrm{idf_i} = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

avec $|D|$ : nombre total de documents dans le corpus
et $|\{d_j : t_i \in d_j\}|$ : nombre de documents où le terme $t_i$ apparaît (c'est-à-dire $n_{i,j} \neq 0$).

Au final la mesure est la suivante :

$$\mathrm{tfidf_{i,j}} = \mathrm{tf_{i,j}} \cdot \mathrm{idf_i}$$

En 1998, [Brin & Page 1998] proposent l'algorithme de classement *PageRank* qui leur vaudra les débuts d'une renommée inconditionnelle. On passe alors à une deuxième génération où le classement et la pertinence des résultats prévalent sur la simple correspondance booléenne que l'on avait précédemment.

Enfin, aujourd'hui, beaucoup de travaux se concentrent sur l'amélioration de l'interprétation des intentions de l'utilisateur. Que ce soit par le biais d'une interaction avec lui, d'une analyse de son profil ou bien par l'analyse et l'enrichissement de sa propre requête, le moteur de recherche cherche à intégrer une connaissance personnalisée de chacun de ses internautes, pour que ceux-ci aient l'impression qu'ils sont au moins aussi bien compris que par un interlocuteur humain.

Comme nous l'avons vu en introduction, une des principales problématiques de la recherche d'information provient du phénomène de synonymie. Les documents correspondant à ce qu'un utilisateur cherche peuvent contenir des termes différents de ceux utilisés dans la requête. Pour pallier ce manque d'information dans la requête exprimée par l'utilisateur en des termes symboliques, certains systèmes proposent d'étendre la requête avec de nouveaux termes en relation de synonymie ou d'hyperonymie avec les termes de la requête.

Cependant, l'usage de synonymes ou d'hyperonymes des mots de la requête peut s'avérer complexe car certains termes sont polysémiques et ont des synonymes différents en fonction du sens dans lequel ils sont employés.

Ainsi, la reformulation de requête a donc pour double objectif de subvenir aux besoins d'élargissement de vocabulaire et de restriction de sens.

L'extension de la requête peut être réalisée de façon automatique et transparente pour l'utilisateur, auquel cas le système doit choisir lui-même quels termes ajouter à la requête et quels poids leur donner.

Elle peut aussi se présenter sous forme de suggestions à l'utilisateur qui choisira lui-même quels termes ou critères ajouter à sa requête ou du moins donnera des informations complémentaires au système. Cette seconde stratégie s'avère nécessaire lorsque la requête ne contient pas suffisamment d'information pour être désambiguïsée automatiquement. C'est le cas notamment lorsque la requête ne contient qu'un seul terme mais aussi dans des cas particuliers comme pour l'exemple précédemment introduit de la requête *école, port, voile*.

Nous ne traiterons pas ici le cas de l'analyse du profil utilisateur car ceci sort du cadre de notre étude. En revanche nous présentons les dernières avancées sur l'analyse et l'enrichissement de la requête utilisateur, qui constitue une part essentielle de l'interprétation des intentions de l'utilisateur.

Dans le but de proposer de nouveaux termes pour l'extension de la requête, on peut faire usage de ressources calculées sur la collection de documents elle-même ou bien utiliser des ressources complètement externes à la collection de documents dans laquelle la recherche est effectuée, ou encore utiliser une combinaison de ces informations.

#### 2.5.2.1 ...par l'analyse des résultats (*Relevance feedback*)

Avant de chercher à interpréter eux-mêmes les intentions des utilisateurs, les premiers systèmes ont d'abord proposé une plus grande interaction avec l'utilisateur afin que celui-ci ne soit plus limité à un champ texte à remplir de mots-clés. Ainsi que le démontre [Salton & Buckley 1990], le simple fait d'interroger l'utilisateur quant à la satisfaction qu'il a des résultats qui lui sont présentés permet d'améliorer les résultats de la recherche. En effet le moteur peut être ré-interrogé avec une seconde requête issue de cette interaction avec l'utilisateur. Ce principe appelé *Relevance Feedback* est décliné sous différentes méthodes depuis les débuts des travaux sur le sujet et en particulier ceux présentés dans [Rocchio 1971].

On parle également de *pseudo-relevance feedback* quand les documents jugés pertinents sont automatiquement sélectionnés par le système comme étant les premiers du classement de réponse du système. Dans ce cas-là, le système relance immédiatement une autre requête et cette analyse est complètement transparente pour l'utilisateur. On retrouve les fondements de tels systèmes dans [Attar & Fraenkel 1977] et plus tard dans [Buckley *et al.* 1995].

[Crabtree *et al.* 2007] propose une approche assez novatrice pour l'extension automatique de requêtes. La méthode se découpe en trois étapes. La première consiste à identifier quels sont les différents *aspects* de la requête en exploitant les fréquences des combinaisons de mots sur un corpus distinct. Le deuxième étape identifie quels sont les aspects sous-représentés dans la collection de documents retournés en premier lieu par le moteur. Il s'agit de déterminer des listes de vocabulaire lié aux différentes aspects et d'analyser les documents retournés pour quantifier la présence du vocabulaire de chaque aspect. Si la présence du vocabulaire lié à un des aspects est trop faible, on passe à la dernière étape. Cette dernière étape reformule la requête initiale avec chacun des meilleurs mots du vocabulaire associé à l'aspect sous-représenté et conserve la requête reformulée produisant la meilleure somme des scores de représentation des aspects initialement détectés. Le système réitère jusqu'à ce qu'il considère qu'il n'y a plus d'aspect sous-représenté ou jusqu'à ce qu'il n'y ait plus de termes de reformulation.

Pour les 10 requêtes du jeu d'évaluation défini (requêtes dites difficiles), les résultats rapportés pour ce système sont meilleurs que les autres méthodes plus classiques proposées pour comparaison. La requête est reformulée 6 fois sur 10 sans jamais détériorer la précision d'origine, ce qui est assez rare pour être signalé. Les 4 fois où la requête n'a pas été reformulée, il s'avère que la précision de la requête non reformulée était déjà relativement correcte ($\tilde{}$70%).

On retrouve dans les travaux d'[Anick *et al.* 2008] une problématique similaire. En effet, bien que le système ait été conçu pour inférer de nouveaux concepts clés (pouvant contenir plusieurs mots) à vendre aux annonceurs publicitaires à partir de leurs mots-clés, la méthode semble être utilisable pour générer des termes de reformulation. Il s'agit d'une variation du *Relevance Feedback* dans laquelle le système remonte pour chaque document un vecteur de dimension limitée représentatif des

syntagmes les plus saillants. Les vecteurs des meilleurs documents retournés pour la requête initiale sont fusionnés pour donner un vecteur global. Les syntagmes de plus hauts scores dans ce vecteur global sont conservés. De nouveaux syntagmes sont générés par combinaison des termes les plus significatifs de la requête et des syntagmes du vecteur global qui sont prédéterminés comme étant des syntagmes rarement utilisés seuls dans une requête (par une pré-analyse des logs de requête). Les syntagmes conservés et les nouveaux syntagmes générés sont ensuite utilisés pour interroger une nouvelle fois l'index du moteur. Les similarités cosinus entre les vecteurs globaux générés par ces syntagmes et le vecteur global généré par la requête initiale produisent un indice de confiance des nouveaux syntagmes en tant que nouveaux concepts clés. Nous pensons que ces concepts clés pourraient être utilisés comme termes de raffinement de la requête.

On peut également analyser l'ensemble des documents retournés par la première requête et proposer à l'utilisateur de raffiner sa recherche en fonction de différents aspects trouvés dans la collection de documents. C'est ce que fait parmi d'autres le moteur de recherche Exalead (*http://www.exalead.com*) en proposant par exemple de rechercher par termes associés, types de site (blog, forum...), types de fichier, langue, pays... C'est ce qu'on appelle la navigation par facettes ou par sérendipidité.

En amont des deux types d'interactions que nous venons de voir, les systèmes peuvent aussi s'intéresser directement à la requête de l'utilisateur et l'analyser pour chercher à interpréter le plus de choses possibles de façon autonome.

#### 2.5.2.2 ...par les ressources manuelles

[Voorhees 1994] exploite les relations lexico-sémantiques présentes dans WordNet pour étendre les requêtes utilisées sur des corpus issus de la campagne d'évaluation TREC. Les relations utilisées sont celles de synonymie, hyponymie, hyperonymie et toutes relations confondues. La sélection des termes à étendre dépend de leur fréquence d'occurrence dans la collection de documents. Les résultats montrent que pour les requêtes longues l'apport de l'extension de requêtes n'est pas significatif. En effet, les requêtes contiennent des descriptions suffisamment complètes de ce que l'utilisateur peut espérer trouver dans les documents réponses. En revanche, il apparaît que l'extension de requête peut améliorer les résultats des requêtes courtes lorsque le vocabulaire utilisé dans les documents ne contient pas les termes exacts utilisés dans les requêtes courtes. Les résultats des expériences menées montrent également la nécessité de sous-pondérer les termes d'extension pour conserver une bonne pertinence des résultats.

[Navigli & Velardi 2003] exploitent les relations sémantiques de WordNet pour les mots monosémiques et les mots polysémiques désambiguïsés. L'étude détaille la reformulation à l'aide de cinq relations différentes dont une exploite les noeuds communs aux réseaux sémantiques d'au moins deux des termes de la requête. Généralement, il s'avère que la relation d'hyperonymie n'est pas un très bon

critère d'extension car les termes additionnels obtenus sont trop généraux. En revanche, les termes appartenant aux définitions des synsets et appartenant donc au domaine sémantique des mots de la requête apportent des résultats nettement plus intéressants.

Le lecteur pourra se référer à l'état de l'art établi par [Bhogal *et al.* 2007] pour une étude plus complète des techniques d'extension de requêtes utilisant des ontologies. Les auteurs distinguent les approches utilisant des connaissances dépendantes du corpus traité (exploitant généralement le *relevance feedback*) de celles utilisant des connaissances purement indépendantes (utilisant WordNet ou des ontologies dédiées). Ils identifient également différents facteurs de succès ainsi que différentes pistes de recherche importantes pour les travaux à venir.

#### 2.5.2.3 ...par les ressources automatiques

Après un certain nombre de tentatives infructueuses d'extension de requêtes effectuées à partir de ressources acquises automatiquement (cf. [Peat & Willett 1991]) où chaque terme de la requête était étendu de façon indépendante, les stratégies proposées s'orientent vers une prise en compte globale de la requête à étendre où chacun des termes de la requête participe d'une façon ou d'une autre au choix des termes d'expansion. C'est le cas des méthodes proposées par [Qiu & Frei 1993] ou encore [Jing & Croft 1994] et beaucoup d'autres par la suite.

[Qiu & Frei 1993] proposent une méthode où tous les termes du vocabulaire sont caractérisés dans le *Document Space Model* par leur appartenance aux différents documents de la collection. Ainsi la similarité entre une requête et un terme candidat à son extension est calculée par une somme pondérée de la similarité de chacun des termes de la requête avec le terme candidat. La pondération est donnée par la pondération de l'importance des termes d'origine de la requête. De cette méthode résulte une amélioration de la précision des résultats pour chacune des trois collections utilisées dans l'évaluation.

Dans les travaux de [Xu & Croft 2000], la méthode se fonde sur la combinaison d'approches globale (*PhraseFinder*) et locale (*Relevance Feedback*). Ils montrent que la pertinence des systèmes fondés sur des approches locales dépend très fortement de la pertinence des éléments les mieux classés et pris en compte dans le calcul de termes d'extension. La méthode combinée proposée *Local Context Analysis* présente de meilleurs résultats car la méthode dispose d'une meilleure métrique pour le choix des termes d'extension.

Une autre approche plus récente calcule néanmoins les termes d'extension indépendamment pour chaque mot de la requête à condition d'y ajouter indirectement des contraintes globales. C'est le cas des travaux de [Vechtomova *et al.* 2003], qui montrent par une première expérience que les cooccurrents globaux (calculés sur l'ensemble de la collection de documents indexés) ne sont pas suffisamment discriminants pour donner une reformulation pertinente. Avec la deuxième méthode proposée (*Local collocation analysis*), les scores (Mutual Information et score Z) de cooccurrence

des termes sont calculés sur un sous-ensemble de la collection, correspondant aux meilleurs résultats renvoyés pour une requête donnée (variation du *pseudo-local feedback*). Cette seconde méthode améliore les requêtes courtes mais n'apporte pas de gain significatif par rapport à la méthode *Okapi Relevance Feedback* de [Sparck Jones *et al.* 2000].

Les auteurs concluent ces expériences en affirmant que les termes d'expansion ne peuvent être calculés sur la collection entière si ceux-ci sont sélectionnés indépendamment pour chaque terme de la requête : une solution serait de les sélectionner en fonction de la requête entière comme dans [Qiu & Frei 1993] ou [Jing & Croft 1994], mais ils peuvent aussi être calculés sur les documents les plus pertinents par rapport à la requête afin de filtrer les contextes d'usages ne correspondant pas à ceux de la requête (prise en compte locale).

Dans une série d'études ultérieures, [Vechtomova *et al.* 2006] et [Vechtomova & Karamuftuoglu 2007] émettent et valident l'hypothèse selon laquelle les contextes des mots de la requête trouvés dans les documents pertinents sont plus lexicalement cohérents que les contextes des mots trouvés dans les documents non pertinents. En utilisant les termes contenus dans l'intersection des contextes issus du *relevance feedback* comme termes d'extension de requêtes, ils parviennent à améliorer significativement la pertinence des résultats.

Une approche intéressante de désambiguïsation non explicite est proposée par [Zhao *et al.* 2006]. En effet, le système est abstrait de tout référentiel de sens. Chaque terme de la requête participe à la sélection des termes d'extension en fonction d'une mesure de similarité sémantique (calculée ici sur le corpus de documents lui-même) qui les lie aux termes d'extension candidats. Les résultats sont nettement améliorés à la fois par rapport à la requête sans extension et par rapport à la requête étendue indépendamment par terme sans tenir compte du contexte (non désambiguïsée).

#### 2.5.2.4 Bilan

Les méthodes recensées ici font plus souvent appel à une désambiguïsation indirecte notamment via le *relevance feedback* qu'à une désambiguïsation standard de la requête. Les résultats dans cette direction sont très prometteurs.

Reprenons maintenant les résultats de l'évaluation de CLEF 2009 *Robust track* ([Agirre *et al.* 2009], cf. section 2.5.1) à laquelle participaient certains systèmes utilisant l'extension de requête après désambiguïsation.

Indépendamment de la stratégie utilisée (indexation par sens et/ou reformulation de requêtes) et de l'aspect monolingue ou crosslingue, ces résultats montrent que certains participants ont effectivement des résultats légèrement meilleurs avec leur système utilisant les informations de désambiguïsation, mais que sur l'ensemble de la tâche, le système fournissant les meilleurs résultats n'exploite pas ces données. La question reste donc posée : la désambiguïsation des corpus et des requêtes, si performante soit-elle, améliore-t-elle la pertinence des résultats d'un système de recherche

de documents ?

Nous remarquons néanmoins que peu d'études ont envisagé une désambiguïsation par l'utilisation de ressources automatiques externes, notamment l'utilisation de clusters de sens et c'est vers cette stratégie que nos travaux s'orientent.

### 2.5.3 Les systèmes de réponses à des questions (Q/R))

Parmi les différents domaines où l'analyse sémantique est de grand intérêt, nous nous intéressons aux systèmes de réponses à des questions en langage naturel. L'origine de tels systèmes remonte aux années 60 où il s'agissait d'interfacer des systèmes experts avec un module de traitement automatique de la langue. Ils représentent maintenant une tâche connexe des systèmes de dialogue homme-machine ([Vilnat 2005]), les avancées des uns pouvant faire progresser les autres.

#### 2.5.3.1 Les systèmes de Q/R en Recherche d'Information

Aujourd'hui, les systèmes de réponses aux questions posées en langage naturel sont une variante des moteurs de recherche traditionnels. De la même façon, les systèmes de Q/R cherchent parmi une collection de documents à la différence que l'utilisateur n'utilise plus exclusivement des mots clés. Il entre sa question telle qu'il l'aurait posée à un humain. Il n'attend en retour non pas une collection de documents comme c'est le cas dans les moteurs traditionnels, mais une réponse à une interrogation précise. Les systèmes permettant de répondre à de telles questions sont généralement constitués de la même structure séquentielle. Les quatre grandes phases que l'on peut identifier sont les suivantes :

Tous les systèmes ne sont pas dotés d'un module d'*analyse et indexation du corpus* mais c'est généralement le cas pour les systèmes les plus performants, à la fois pour la vitesse d'exécution et pour la pertinence des résultats. Cette étape est effectuée en amont, il n'est pas nécessaire de la reproduire à chaque nouvelle question contrairement aux étapes suivantes.

L'*analyse de la requête* va permettre au système d'effectuer différents traitements dont les plus répandus sont la distinction entre mots vides (*stop words*), mots pleins et éventuellement détection d'*expressions multimots*†, ainsi que le typage de la question (qu'est-ce qui est attendu comme réponse ? est-ce une personne, un nombre, une raison, une liste, une définition... ?). Cette étape permet aussi l'extension de la requête à l'aide de simples synonymes ou d'algorithmes plus complexes (voir section 2.5.2). Enfin, c'est aussi à ce moment-là que certains systèmes effectuent une analyse sémantique plus poussée allant de la détection d'entités nommées à l'extraction de prédicats logiques, en passant par la désambiguïsation de sens, la traduction de termes ou encore la correspondance avec les classes d'une ontologie.

Une fois l'analyse de la requête terminée, le système *reformule la question d'origine* en une requête *ad hoc* et interroge son index pour récupérer et classer un certain nombre de documents ou de

passages dans lesquels la probabilité de trouver la réponse est forte.

Enfin, il s'agit pour le système d'*extraire la ou les bonnes réponses*. Pour les questions factuelles, il s'agit quasi-systématiquement d'un syntagme seul. Une autre tâche d'importance est d'identifier la ou les morceaux de passages justifiant la réponse donnée afin que l'utilisateur sache sans hésitation pourquoi la réponse lui est faite. Il est aussi parfois important de pouvoir donner un score de confiance à ses réponses afin que l'utilisateur sache à quoi s'en tenir.

Pour un détail plus précis de tels systèmes, on pourra se référer à [Chalendar (de) *et al.* 2002], [Amaral *et al.* 2004], [Laurent & Séguéla 2005], [Moldovan *et al.* 2007].

Le meilleur système de question réponse de la campagne TREC 2007 ([Dang *et al.* 2007]) est celui de [Moldovan *et al.* 2007]. Les modules les plus spécifiques de ce système sont un analyseur temporel et un moteur de raisonnement logique exploitant les *glosses* de WordNet. Les évaluations TREC menées les années suivantes ne sont pas rapportées dans ce manuscrit car la tâche s'est orientée vers la prise en compte des opinions. Nous limiterons notre étude aux questions factuelles (*qui, quand, où, ...*) et complexes ou descriptives (*pourquoi, comment, définitions...*).

#### 2.5.3.2 L'annotation sémantique de rôles dans les systèmes de Q/R

Les premiers travaux utilisant l'annotation en rôles sémantiques dans des tâches plus appliquées concernent l'extraction d'information. Cette tâche partage des problématiques communes avec les systèmes de Q/R. Le lecteur pourra se référer aux travaux de [Moschitti *et al.* 2003] et [Surdeanu *et al.* 2003] pour de plus amples détails.

En ce qui concerne plus directement les systèmes de questions/réponses, [Narayanan & Harabagiu 2004] proposent un modèle complexe pour l'inférence sémantique. Ce modèle exploite à la fois les structures prédicat-arguments de PropBank et les *frames* de FrameNet par une réimplémentation des méthodes proposées respectivement dans [Gildea & Palmer 2002], [Gildea & Jurafsky 2002]. Ces relations sémantiques sont utilisées avec des *topic models* pour générer une structure événementielle. Cette structure événementielle est conçue pour appliquer des règles d'inférence. Les résultats donnés pour l'identification du type de la réponse sur des questions complexes (facteur identifié précédemment par [Moldovan *et al.* 2002] comme étant le plus gros problème des systèmes de Q/R), présentent une amélioration impressionnante par rapport au système sans traitement sémantique. Cet article est le premier à montrer des améliorations significatives suite à l'utilisation de relations sémantiques de type *prédicat-arguments* ou *frames* combinée à une représentation sémantique complexe et à son modèle d'inférence.

[Sun *et al.* 2005] utilisent les relations sémantiques issues d'un analyseur de type PropBank pour identifier un ensemble d'arguments caractérisant une phrase dans le but de la comparer avec la

question. Le type des arguments n'est cependant pas pris en compte. La première raison à cela est le taux d'erreurs encore assez élevé de la tâche de classification des arguments, en particulier sur un corpus bruité tel que le Web. La deuxième raison provient du fait que les arguments Arg-0, Arg-1, Arg-2 ne signifient pas la même chose d'un verbe à un autre, ce qui est gênant quand deux structures prédicatives comparées ne sont pas les mêmes.

Les résultats sont rapportés comme étant meilleurs que ceux de l'année précédente pour les questions factuelles mais l'évaluation ne distingue pas l'apport du module d'extension de la requête de celui du module d'extraction de la réponse sur les documents Web (answers.com).

De la même façon que [Sun *et al.* 2005], [Stenchikova *et al.* 2006] utilisent les relations sémantiques d'un analyseur de type PropBank pour identifier les arguments du prédicat de la question et les soumettre comme requête au système. Cette stratégie s'avère plus performante lorsqu'elle est utilisée en *fallback*, c'est-à-dire lorsque la stratégie *exact match* correspondant à la question mise sous forme affirmative ne retourne aucun résultat. Enfin, les relations sémantiques sont également utilisées pour l'extraction de la réponse où seuls les arguments de la phrase candidate identifiés comme correspondant au type de la question sont annotés comme réponse potentielle. En revanche, l'annotation des pronoms interrogatifs n'étant que très peu précise, les règles de correspondance *pronom interrogatif -> type de l'argument réponse* sont issues d'heuristiques pour les questions *Who, When* et *Where* et d'un classifieur pour les questions *What*. Les résultats calculés sur un ensemble de 190 questions contenant toutes des prédicats (et non le verbe *to be*) montrent une meilleure précision. De plus, l'évaluation utilisateur montre que les réponses du système avec SRL présentent un meilleur taux d'information purement pertinente, et un meilleur taux d'extraits grammaticalement corrects, que les réponses du système sans SRL.

Les travaux de [Kaisser & Webber 2007] montrent des résultats similaires. Les auteurs exploitent cette fois les trois ressources connues en sémantique de rôles, à savoir FrameNet, PropBank et VerbNet. Ils définissent de la même façon un certain nombre de règles pour déterminer le type de la réponse cherchée et proposent ensuite deux stratégies de recherche de la réponse. La première produit plusieurs requêtes structurées, en langage naturel, en employant toutes les combinaisons de voix et de temps compatibles avec ceux de la question, et en laissant un espace vide à l'endroit où la réponse potentielle se trouverait. Des adaptations particulières sont proposées pour bénéficier de toute la richesse de FrameNet (regroupement de plusieurs unités lexicales sous une même frame et relations entre les frames) et produire ainsi des requêtes supplémentaires correspondant à des paraphrases des requêtes initialement générées. La deuxième stratégie est moins stricte, tous les mots pleins de la question sont envoyés en une seule requête. Une heuristique de score est ensuite définie par rapport à la présence des différents termes de la question et rôles du verbe dans la phrase candidate à la réponse. Les apports respectifs des trois sources et des deux stratégies sont clairement étudiés par les expériences mises en œuvre. Enfin, sur le corpus de questions-réponses de TREC 2002 duquel ont également été retirées les questions dont le verbe principal était *to be*, le gain en

précision par rapport au système des auteurs sans ce module (en 2006) est quantifié à 21%.

[Moschitti *et al.* 2007] utilisent les rôles sémantiques (structure prédicat-arguments) comme nouveaux traits dans leur apprentissage supervisé. Ils montrent que ceux-ci n'apportent aucun gain concernant la classification des questions. Ceci s'explique en partie par le fait que la moitié des questions comportent le verbe *to be* en tant que verbe principal et qu'il n'existe pas de structure prédicative pour ce verbe. En revanche, leur utilisation révèle un impact très positif sur la classification des réponses (est-ce qu'une phrase ou un passage répond ou non à une question ?) et leur ré-ordonnancement (le système atteint un MRR de 81%). L'évaluation est effectuée sur des questions de type *description* dites difficiles.

Outre le fait de décrire et d'évaluer une nouvelle méthode d'apprentissage des rôles sémantiques [Moreda *et al.* 2007] proposent une approche préliminaire à l'utilisation des rôles sémantique dans les systèmes de Q/R. L'annotation en rôles sémantiques de type PropBank se fait en trois étapes : désambiguïsation du sens des verbes, détection des frontières des arguments, et identification des rôles sémantiques. Chacune des étapes est traitée à l'aide de 2 méthodes statistiques (*Maximum Entropy*/*Memory Based Learning TiMBL*) pour lesquelles les auteurs appliquent un algorithme itératif de sélection de traits distinctement pour chacune des trois étapes. Les résultats sont évalués en validation croisée et c'est la méthode TiMBL qui donne les meilleurs résultats avec une F-mesure de 76%. La partie concernant l'utilisation de ces rôles pour les systèmes de Q/R est uniquement descriptive. L'algorithme sélectionne d'abord les phrases pertinentes pour la question, c'est-à-dire les phrases contenant un verbe appartenant à la liste des troponymes et synonymes du verbe de la question. Ces phrases sont annotées à l'aide du système de SRL précédemment décrit. Les rôles PropBank sont ensuite appariés à un ensemble plus descriptif de rôles (heuristiques spécifiques à la phrase dans les cas d'ambiguïté dans le mapping). Enfin les auteurs définissent une topologie de questions auxquelles ils associent un ensemble de rôles réponses possibles. Seules les phrases contenant un argument remplissant ces rôles sont conservées pour la suite des traitements. Aucune étude quantitative n'est menée sur cette stratégie.

Dans la suite des travaux de sa thèse, Paloma Moreda analyse dans [Moreda *et al.* 2008] les différents systèmes de Q/R utilisant les rôles sémantiques et en vient à la conclusion que les rôles sémantiques ne peuvent pas jouer un très grand rôle dans la détection du type de la question (du fait des faibles résultats des systèmes d'annotation supervisés sur les phrases interrogatives), mais qu'en revanche la détection de ces rôles pourrait se révéler très intéressante dans l'extraction de la réponse. Afin de valider cette hypothèse, les auteurs analysent les résultats de trois systèmes différents, dont un est une implémentation du système de [Pizzato & Mollá-Aliod 2005] fondé uniquement sur l'utilisation des Entités Nommées (EN) pour l'extraction de la réponse ; le deuxième utilise les règles de mapping manuelles entre types de question et rôles sémantiques (définies dans [Moreda

*et al.* 2007]) et le troisième système utilise des patrons sémantiques appris automatiquement dans le but de traiter le problème des *lieus* qui apparaissent comme arguments principaux ambigus comme dans la phrase *Marie va au parc.* et non comme Arg-M comme dans la phrase *Marie joue dans le parc.* Les questions sont restreintes au type *lieu* pour réduire le biais dû aux différentes questions. L'évaluation est menée sur deux types de questions *lieu*, celles dans lesquelles la question contient généralement une entité nommée et attend une réponse entité nommée comme *Quelle est la plus grande ville d'Allemagne ?* et celles ne contenant que des noms communs comme *Où se trouve le pancréas ? dans l'abdomen.* Les résultats confirment bien l'hypothèse intuitive, le premier système (EN) résout très bien les questions Entités Nommées (MRR de 87% contre 52% et 58% pour les systèmes à base de SRL) mais n'est pas du tout adapté aux questions noms communs (MRR de 13%). En revanche, les systèmes à base de SRL résolvent mieux ce type de question (MRR de 41%, précision de 95% !)

L'étude menée par [Shen & Lapata 2007] concerne l'apport d'une analyse sémantique fondée sur FrameNet aux systèmes de Q/R. Elle répond aux trois questions suivantes :

- Comment la non-exhaustivité de FrameNet influence-t-elle la performance d'une analyse sémantique pour les systèmes de Q/R ? et plus exactement, quelle proportion de questions n'obtient pas de réponse du fait de l'absence de prédicat dans la bonne *frame* ou de l'absence d'un exemple annoté de ce prédicat ?
- Est-ce que toutes les réponses correspondant à une question peuvent être mises en relation par une analyse fondée sur FrameNet ? Autrement dit, combien trouve-t-on de cas où l'analyse de la question ne peut pas être mise en correspondance avec l'analyse de la phrase réponse en raison de l'absence de relation entre les deux *frames* détectées ?
- Enfin, la méthode de SRL proposée étant spécifiquement conçue pour les systèmes de Q/R, apporte-t-elle un gain par rapport à une méthode purement syntaxique ou bien sémantique traditionnelle ?

Sur les questions étudiées manuellement (questions factuelles TREC 2002 à 2005), environ 29% n'ont pas de *frame* correcte correspondant au prédicat de la question, 7% n'ont pas d'annotation (correspondant à la catégorie grammaticale du prédicat pour la frame associée) dans le corpus d'apprentissage, 41% des analyses effectuées sur les questions et phrases réponses ne peuvent pas être mises en correspondance et enfin seulement 32% des questions peuvent effectivement être mises en relation avec leur phrase réponse par une annotation sémantique de type FrameNet. La méthode de SRL proposée par les auteurs améliore en effet significativement les résultats sur les 32% pouvant être mis en relation. Les auteurs prônent aussi l'utilisation d'une approche purement syntaxique complémentaire dans le cas où aucune réponse n'est trouvée par le système SRL afin de pallier la non-exhaustivité de FrameNet. Ils montrent également des résultats significatifs sur l'ensemble des données.

Du fait de la faible performance en vitesse des systèmes de SRL, [Pizzato & Mollá 2008] proposent une méthode d'analyse sémantique *de surface* pour indexer les documents d'un système de question-réponse. Les résultats obtenus sont moins bons que ceux obtenus avec une véritable annotation en rôles sémantiques mais néanmoins meilleurs qu'un simple système *sac de mots*. En revanche, le gain en vitesse est très important.

#### 2.5.3.3 Le raisonnement et le SRL dans les systèmes de Q/R

[Salvo Braz (de) *et al.* 2005] modélisent la tâche de Questions/Réponses comme étant un problème de *Recognising Textual Entailment (RTE)* dans lequel le système doit déterminer si une phrase contenant potentiellement une réponse implique la question donnée en entrée. Si c'est bien le cas, alors la réponse attendue est contenue dans la phrase candidate. Par exemple la phrase *a) Jean a acheté le livre hier en ville* répond à la question *Qui a acheté le livre ?* puisque *a* implique *XXX a acheté le livre*. Dans ce cadre là, les auteurs étudient leur modèle de raisonnement fondé sur une logique de description *Extended Feature Description Logic (EFDL)* modélisant trois niveaux d'abstraction : le lexique au niveau des mots, la syntaxe au niveau des syntagmes, la sémantique au niveau des rôles de type PropBank, ainsi que des règles de réécriture conservant le sens d'une représentation à une autre. Les représentations des graphes conceptuels (équivalents à la logique de description utilisée) sont ainsi générées par un étiqueteur, un analyseur syntaxique en arbre, un analyseur sémantique de type PropBank, un détecteur d'entités nommées, un solveur de coréférences nominales et pronominales, un tokenizeur et un lemmatiseur. Les règles de réécriture lexicales sont filtrées à partir du corpus de paraphrases généré par [Lin & Pantel 2001], et un certain nombre de règles de réécriture non lexicales sont écrites à la main, produisant au total un ensemble de plus de 300 règles d'inférence. Les auteurs étudient 3 niveaux d'analyse (lexical, lexical+SRL, lexical+SRL+modifieurs(adjectifs, adverbes, certains déterminants)) ainsi que l'impact de différents modules supplémentaires : analyse verbale (modalités, polarités, temps, voix), analyse du discours dans des constructions de type *X dit que Y VP* ou *X dit que Y, qui VP1-passé, VP2* dans lesquelles VP et VP2 n'ont qu'une confiance relative, tandis que VP1 serait assumé comme étant vrai, analyse des quantifieurs (tous, beaucoup, plusieurs, un, aucun...). L'évaluation de ces *entailments* est faite sur le corpus fourni par Xerox-PARC[28] spécifiquement conçu pour tester différents cas d'inférences linguistiques dans le cadre de question/réponses. Le corpus contient 76 paires de (questions reformulées à l'affirmative, réponses). Les systèmes proposés montrent des résultats intéressants en particulier lorsque l'analyse SRL est effectuée sans erreurs. Chacun des modules présentés apporte une amélioration significative pour atteindre une précision maximum de 83%.

Les campagnes d'évaluation du challenge PASCAL RTE proposent pour la cinquième fois en 2009 une tâche de reconnaissance de déduction textuelle ([Bentivogli *et al.* 2009]). Un texte source est

28. `http://l2r.cs.uiuc.edu/~cogcomp`

associé à une hypothèse et les systèmes doivent déterminer si le texte implique l'hypothèse, s'ils sont contradictoires ou s'ils sont indépendants. Une partie de ces paires est construite à partir de corpus de questions-réponses issus des campagnes CLEF et TREC. Le meilleur des systèmes candidats à la dernière évaluation obtient une précision de 68% et la meilleure précision atteint les 74% dans le cas où on ne distingue pas la contradiction de l'indépendance des assertions. Parmi les 20 participants à l'évaluation, seuls 3 utilisent FrameNet et 1 utilise PropBank. Par ailleurs les tests effectués sans ces ressources ne montrent pas l'aspect fondamental de leur usage pour la déduction textuelle.

#### 2.5.3.4 Bilan

Les premiers résultats d'intégration de module de SRL dans des systèmes de questions/réponses sont très prometteurs. A l'issue de nos travaux, nous souhaitons mettre en œuvre notre propre système de SRL dans un système de Q/R existant afin de valider son efficacité et de démontrer son intérêt applicatif. Ceci n'a à ce jour jamais été possible pour les systèmes de langue française.

### 2.5.4 Conclusions

Nous venons de voir où en sont les systèmes de recherche d'information quant à la prise en compte d'information sémantique de haut niveau. Nous tenterons en conclusion de nos travaux de répondre aux questions qui en découlent.

L'indexation de mots par sens n'est pas encore d'un intérêt reconnu bien que beaucoup croient en son potentiel. Est-il possible de mettre en œuvre un système faisant bon usage de cette indexation par sens ?

L'extension de requête exploite par nature de l'information issue de ressources et d'analyse sémantiques. La désambiguïsation lexicale est-elle utilisée de façon optimale dans les approches existantes ? L'annotation sémantique de rôles peut-elle participer à ce processus d'extension ?

Enfin, la plupart des systèmes de question-réponse exploitent une analyse sémantique plus ou moins formalisée (typage des questions, désambiguïsation, reconnaissance d'entités nommées, de dates, de types de réponses...). L'usage des rôles sémantiques est néanmoins loin d'être généralisé à tous ces systèmes. L'annotation en rôles sémantiques améliorerait-elle les systèmes existants français n'en faisant pas usage ?

## 2.6 Bilan

Nous venons de passer en revue diverses techniques de fouille de données spécifiques au traitement des données multi-représentées et de grandes dimensions. Nous avons également fourni un recensement des principales ressources sémantiques manuelles et automatiques mises à disposition de la communauté et décrit un panel de différentes méthodes et résultats de traduction et d'enrichissement de ces ressources. En outre, nous avons analysé et exposé les principales approches de

désambiguïsation lexicale et d'annotation sémantique de rôles. Pour finir, nous avons mis en perspective ces tâches d'analyse sémantique dans le cadre plus applicatif de la recherche d'information.

Après avoir dressé ce panorama du domaine de la sémantique, que nous avons souhaité suffisamment complet pour cerner le contexte de l'ensemble de nos problématiques, nous abordons maintenant la première tâche que nous avons défini en analyse sémantique : la désambiguïsation lexicale.

# Deuxième partie

# Désambiguïsation lexicale

# Chapitre 3

# Création de ressources pour la désambiguïsation lexicale

Quelles que soient les modalités de la désambiguïsation lexicale, celle-ci se fait toujours par référence à une définition de sens. Dans le cas d'une désambiguïsation automatique, il peut être judicieux de référer à une énumération de sens qui permettra l'apposition d'une étiquette. Cette étiquette pourra ainsi être réutilisable par la suite plus facilement qu'une paraphrase ou une autre modalité de définition. C'est en tout cas l'approche adoptée par une très grande partie de la communauté étudiant les systèmes de désambiguïsation automatique et dans toutes les campagnes d'évaluation en désambiguïsation lexicale.

Cette approche nous semble également pertinente dans le cadre de notre travail. L'application la plus probable qui sera amenée à utiliser notre module est un moteur de recherche. Celui-ci présentant une interface graphique, il pourra alors être demandé à l'utilisateur de choisir un ou plusieurs sens au(x)quel(s) il fait référence. Le fait de référer au sens par une étiquette est un paradigme très simple de définition du sens pour un usager non expert, il lui suffit de savoir si oui ou non l'étiquette correspond à ce qu'il cherche.

Nous sommes donc amenée à nous pencher sur la question du choix du répertoire de sens à utiliser. Dans les dernières campagnes d'évaluation de langue anglaise, c'est la ressource WordNet qui a été utilisée et ceci malgré les observations sur la parfois trop grande finesse de sa granularité. WordNet est également utilisé dans un très grand nombre d'autre tâches du traitement automatique des langues. Cependant les équivalents français existants (voir section 2.3) n'égalent pas encore les qualités du WordNet anglais. C'est pourquoi nous proposons dans la première partie de ce chapitre une nouvelle méthode de traduction du WordNet anglais vers le français.

Depuis quelques années, il existe aussi une nouvelle tendance qui vise à construire des répertoires de sens uniquement par l'analyse automatique de corpus. Ces approches consistent à appliquer des méthodes de clustering aux mots du vocabulaire dans le but d'assembler les mots en groupe définissant des sens. Elles permettent à la fois d'obtenir une plus grande couverture du vocabulaire, de pouvoir s'adapter à un domaine ou à une langue particulière et de définir le sens de façon plus continue. Nous proposons dans la deuxième partie de ce chapitre de pratiquer ce type d'induction de sens de mots, en adaptant deux méthodes existantes afin qu'elles puissent exploiter de multiples représentations sémantiques issues d'une analyse de corpus à grande échelle.

## 3.1 Transcription au français d'un réseau lexical constitué manuellement

Il existe déjà plusieurs tentatives de constitution de WordNet pour le français telles que celles développées dans le cadre du projet EuroWordNet ([Vossen 1998]) ou comme la ressource Wolf développée par [Sagot & Fišer 2008], mais aussi pour d'autres langues comme les travaux de [Barbu & Barbu Mititelu 2005] par exemple. Le point le plus délicat de ces transferts d'une langue à une autre réside dans la traduction des mots source polysémiques, et c'est sur ce point particulier que nous souhaitons proposer une approche originale. L'idée principale est d'exploiter les propriétés des distributions des contextes syntaxiques des mots dans un grand corpus afin de caractériser les relations sémantiques présentes dans WordNet. Nous restreignons notre étude aux noms, catégorie la plus représentée dans WordNet.

L'évaluation de ce type de travaux est difficile puisqu'il n'existe pas par définition de vérité terrain sur laquelle s'appuyer. Nous décidons de faire reposer l'évaluation de notre travail d'une part sur une comparaison avec une des précédentes tentatives de constitution d'un WordNet français (Wolf), et d'évaluer manuellement les traductions présentes dans nos résultats mais pas dans Wolf.

WordNet [1]. En effet, les traductions données par un dictionnaire ne correspondent pas forcément à tous les synsets d'un même terme, il s'agit de déterminer la ou les traduction(s) adaptée(s) à chaque synset.

Cette section se divise en trois parties assez classiques. Nous décrivons dans un premier temps l'approche proposée. Nous présentons ensuite la méthode d'évaluation adoptée ainsi que les résultats obtenus. Nous terminons par la mise en perspective de ces résultats accompagnée de suggestions complémentaires.

1. Rappelons ici que la structure de WordNet distingue les sens des mots par le regroupement en *synsets*. Un synset correspond à un ensemble de synonymes associé à une définition. Certains synsets sont reliés entre eux par des relations sémantiques.

### 3.1.1 Approche proposée

Nous présentons maintenant le détail de l'implémentation de notre système de traduction automatique de la ressource WordNet.

#### 3.1.1.1 Vue d'ensemble du système

La structure du WordNet original que nous nommerons *PWN* (Princeton WordNet) est tout d'abord reproduite pour la constitution du WordNet de langue cible. Après une phase d'extraction des candidats de traduction, chaque heuristique définie dans la suite de cette section est appliquée de façon itérative, de sorte que le WordNet cible se remplisse petit à petit et qu'à chaque itération, de nouvelles informations viennent rendre possible de nouvelles traductions. Nous ne traitons dans ce travail que des termes et syntagmes nominaux auxquels nous référerons dès lors plus simplement par l'emploi du mot *terme*.

La phase d'extraction consiste à traduire tous les termes associés à un seul synset par toutes les traductions proposées par notre dictionnaire bilingue [2]. Pour les autres termes et chacun de leurs synsets associés, nous conservons toutes les traductions comme termes cible candidats. La désambiguïsation consistera à déterminer quel terme candidat (s'il existe) correspond au sens de chaque synset. La structure du PWN est ainsi conservée : l'appellation synset fait maintenant à la fois référence aux termes source et aux termes cible. Cette étape d'extraction sera notée $E$ dans les résultats présentés plus loin. On parlera de synset instancié pour référer aux synsets auxquels on a assigné au moins un terme cible.

À chaque itération de l'algorithme, on produit autant de nouvelles ressources (ensemble de traductions) que d'heuristiques. Puis, une évaluation automatique [3] est menée pour déterminer quelle heuristique fournit la meilleure amélioration en précision. On élimine les autres et on réitère en utilisant la ressource conservée (Figure 3.1).

Trois des quatre heuristiques que nous proposons exploitent les distributions syntaxiques des mots dans les espaces sémantiques décrits dans la section suivante.

#### 3.1.1.2 Espaces sémantiques

Les espaces sémantiques que nous utilisons dans l'ensemble des thématiques que nous abordons dans notre thèse sont issus des travaux de [Grefenstette 2007].

**Acquisition des espaces** Ces espaces sont calculés à partir d'une analyse en dépendances syntaxiques sur un corpus français issu du Web. Les documents furent obtenus après avoir envoyé 600 000

2. Nous utilisons la concaténation du dictionnaire *SCI-FRAN-EuRADic* (`http://catalog.elra.info/product_info.php?products_id=666&language=fr`) et du Wiktionnaire français (`http://fr.wiktionary.org/wiki/Page_d%27accueil`).

3. L'évaluation considère l'intersection avec Wolf comme vérité-terrain, cf. section suivante

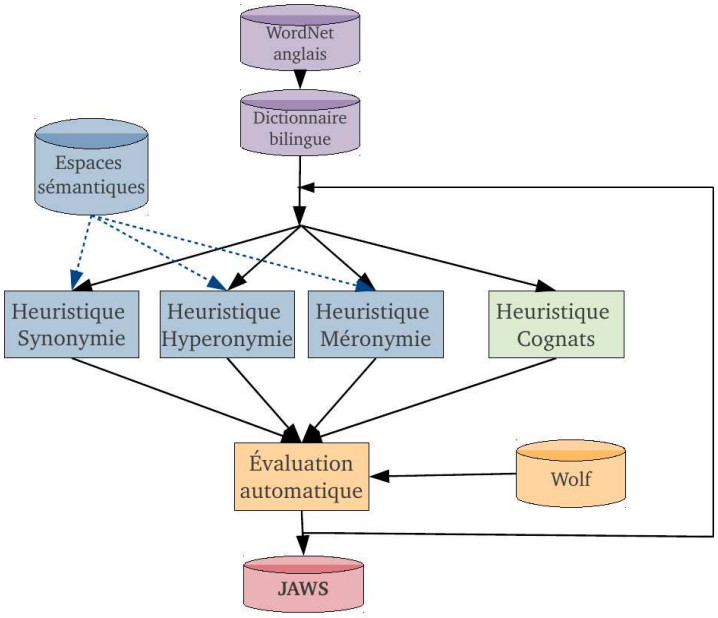



FIGURE 3.1 – Vue générale de notre système

mots d'un dictionnaire comme requêtes sur un moteur de recherche et téléchargé les 100 premiers résultats pour chaque requête ([Grefenstette 2007]).

Au moment où nous effectuons nos travaux, le corpus est constitué de deux millions de documents de pages francophones du Web sur lesquelles a été effectuée une analyse syntaxique par le système d'analyse LIMA du CEA LIST [Besançon & Chalendar (de) 2005]. Ce système effectue une analyse en dépendances syntaxiques de type *sujet_verbe*, *objet_verbe*, *compl_du_nom*, (...). Ces relations sont orientées, par exemple dans le syntagme *traitement des langues*, *langues* apparaît dans le contexte de *traitement* pour la relation *complément du nom* tandis que *traitement* apparaît dans le contexte de *langues* pour la relation *complément du nom* inverse.

L'analyse syntaxique effectuée par LIMA annote le corpus en 34 relations différentes : les relations homo-syntagmatiques relient deux mots à l'intérieur d'un même syntagme non verbal, tandis que les relations hétéro-syntagmatiques relient deux mots qui ne figurent pas dans le même syntagme. La liste des relations identifiées par cet analyseur et utilisées lors de la construction des espaces sémantiques est fournie dans le tableau 3.1. Certaines relations sont détectées par LIMA mais n'ont pas été retenues pour la constitution des espaces pour une raison que nous ignorons. Ces relations ne figurent pas dans le tableau.

Le dictionnaire utilisé pour la constitution des espaces sémantiques est constitué des 68 000 mots les plus fréquents de la langue française. Un espace distinct est associé à chaque relation, enregistrant les fréquences de cooccurrences des 68 000 mots avec les 68 000 contextes ainsi définis

| Relations homo-syntagmatiques | Relations hétéro-syntagmatiques |
|---|---|
| complément du nom | sujet du verbe |
| adjectif épithète pré-nominal | pronom sujet du verbe de la proposition relative |
| adjectif épithète post-nominal | antécédent sujet du verbe de la proposition relative |
| apposition | complément d'objet direct du verbe |
| substantif juxtaposé à un substantif | complément indirect du verbe |
| adverbe modifiant un adjectif | complément circonstanciel du verbe |
| adverbe modifiant un adverbe | attribut de l'objet |
| adverbe modifiant un substantif | attribut du sujet en relation avec le verbe |
| adverbe modifiant un verbe | attribut du sujet en relation avec le sujet |
| composition (adjectif pré-nominal post-nominal ou complément du nom) | complémenteur |
| complément de l'adjectif | |
| complément de l'adverbe | |
| déterminant interrogatif | |
| déterminant numéral cardinal | |
| modificateur de l'adjectif | |
| modificateur du nom | |
| modificateur du verbe | |
| négation | |
| préfixe | |
| relation entre préposition et déterminant interrogatif | |
| relation entre préposition et conjonction de subordination considérée comme un pronom clivé | |
| relation entre préposition et pronom relatif complément d'attribution | |
| relation entre préposition et pronom relatif | |
| relation entre préposition et pronom personnel | |

TABLE 3.1 – Relations de dépendance syntaxiques détectées par LIMA et utilisées pour construire les espaces sémantiques

pour cette relation. On donne ci-dessous dans le tableau 3.2 un extrait des matrices *complément d'objet (COD_V)* et *complément du nom- COMPDUNOM* construites avec les phrases suivantes :

*On étudie actuellement les stratégies de traitement de signaux afin d'obtenir les résultats souhaités. (...) Le responsable d'un traitement de données doit obtenir le consentement préalable des personnes concernées avant toute utilisation de ces données. (...) Il a obtenu un doctorat en Traitement Automatique des Langues.*

L'analyse syntaxique trouve entre autres que *stratégie* est complément d'objet (*COD*) de *étudier* et que *traitement* est complément du nom (*COMPDUNOM*) de *stratégie*, ce qui se traduit dans la matrice du tableau 3.2 par l'incrémentation des cellules correspondantes.

Cette matrice, si petite soit-elle, donne déjà une première intuition de la métaphore géométrique donnée par les espaces sémantique. On retrouve des mots représentés par des vecteurs identiques, comme les mots *signal*, *donnée* et *langue* ou les mots *résultat*, *consentement* et *doctorat*. En analysant un nombre de contextes beaucoup plus important, on parvient à une modélisation beaucoup plus fine d'une telle similarité de sens, fournissant ainsi une distance sémantique quantifiable entre les mots.

| | COD_V | | COMPDUNOM | | | |
|---|---|---|---|---|---|---|
| | étudier | obtenir | stratégie | traitement | consentement | utilisation |
| stratégie | 1 | 0 | 0 | 0 | 0 | 0 |
| résultat | 0 | 1 | 0 | 0 | 0 | 0 |
| consentement | 0 | 1 | 0 | 0 | 0 | 0 |
| doctorat | 0 | 1 | 0 | 0 | 0 | 0 |
| traitement | 0 | 0 | 1 | 0 | 0 | 0 |
| signal | 0 | 0 | 0 | 1 | 0 | 0 |
| donnée | 0 | 0 | 0 | 1 | 0 | 1 |
| personne | 0 | 0 | 0 | 0 | 1 | 0 |
| langue | 0 | 0 | 0 | 1 | 0 | 0 |

TABLE 3.2 – Matrices de cooccurrences syntaxiques

Afin que l'information fournie par tous les mots, y compris les mots rares, puisse être prise en compte, nous travaillons avec des matrices d'*information mutuelle spécifique* calculées à partir des matrices de fréquence. L'information mutuelle spécifique (*Pointwise Mutual Information*) est calculée à partir de la formule 3.1, où $P_i$ est la probabilité d'occurrence du terme décrit par la ligne $i$ dans n'importe quel contexte de la relation donnée, $P_j$ est la probabilité d'occurrence du contexte défini par la colonne $j$, et $P_{i,j}$ est la probabilité de cooccurrence du terme $i$ avec le contexte $j$ pour la relation donnée. L'information mutuelle est positive si la probabilité de cooccurrence des termes $i$ et $j$ est plus grande que la probabilité attendue si ces événements étaient indépendants.

$$PMI(i,j) = log(\frac{P_{i,j}}{P_i * P_j}) \tag{3.1}$$

La même procédure est effectuée pour les relations inverses, ainsi que pour les cooccurrences dans

des fenêtres de taille fixe (5, 10, 20). Au final, on dispose de 71 matrices creuses et carrées de 68000 dimensions. Nous gardons ces matrices séparées, distinguant ainsi 71 espaces sémantiques dans lesquels sont représentées les mêmes données. 34 espaces sont construits à partir des relations syntaxiques énumérées plus haut, 34 autres espaces sont construits à partir des matrices transposées des 34 premières et correspondent à la relation inverse des relations syntaxiques utilisées. Enfin, 3 espaces sont construits à partir de cooccurrences de fenêtre de taille fixe (5, 10 et 20). On verra par la suite que ces 71 espaces conservent des informations différentes et complémentaires.

**Réduction de dimensions : *Locality Sensitive Hashing*** La taille de ces matrices nécessite une réduction du nombre de dimensions afin de travailler sur des matrices de taille raisonnable. Les décompositions en valeurs singulières des méthodes d'Analyse Sémantique Latente (*Latent Semantic Analysis, LSA*) développées par [Landauer & Dumais 1997] devenant assez lourdes sur nos matrices (complexité quadratique), nous nous tournons vers des méthodes de réduction par Hachage Sensible à la Localité (*Locality Sensitive Hashing, LSH*), qui sont plus adaptées à la taille de nos matrices.

[Charikar 2002] définit une famille de hachage sensible à la similarité cosinus. Nous rappelons ici la formule du cosinus approximatif :

$$\cos\theta(\overrightarrow{u}, \overrightarrow{v}) \approx \cos \frac{distance\_de\_Hamming(\overrightarrow{u}, \overrightarrow{v})}{d} * \Pi$$

Pour les détails concernant la définition des familles de hachage et les démonstrations, nous invitons le lecteur à se référer à la section 2.1.2.2 de notre état de l'art.

Cette méthode de réduction nous fournit en sortie autant de signatures (vecteurs de bits) que nous avions de vecteurs en entrée. Nous avons constaté empiriquement par l'analyse manuelle des plus proches voisins obtenus que des signatures de taille 16 384 bits nous donnaient des résultats satisfaisants.

Ayant ainsi présenté les différents espaces que nous utilisons, nous pouvons à présent définir nos heuristiques de désambiguïsation. Celles-ci exploitent les relations sémantiques du Princeton WordNet ainsi que des caractéristiques de distribution des termes cible (français) dans les espaces sémantiques que nous venons de décrire.

#### 3.1.1.3 Heuristiques

**Synonymie** La première heuristique, désignée par $S$ dans les résultats, exploite une mesure de similarité sémantique. Nous utilisons le cosinus approximatif dans les espaces décrits plus haut. Cette mesure permet de trouver des relations proches de la synonymie [Turney 2001].

Soit $t$ le terme source d'un synset. Si le terme $t$ a plusieurs traductions candidates, et que le synset a déjà été instancié (par une traduction non-ambiguë ou une étape précédente de traduction de terme ambigu), alors la traduction choisie est celle la plus proche des termes cible instanciés.

Nous présentons à la figure 3.2 un exemple traité par notre système grâce à cette heuristique. La figure représente le synset dont la définition est *A condensed but memorable saying embodying some important fact of experience that is taken as true by many people*[4]. Le terme anglais *saw* appartient à ce synset et se traduit en français par *dicton* ou *scie*. Ces deux candidats de traduction apparaissent dans l'ellipse jaune.

Le système doit choisir lequel des deux est la meilleure traduction pour ce synset. On voit également sur la figure que le terme *saw* est en relation de synonymie avec trois autres termes anglais *adage*, *byword* et *proverb*. Deux d'entre eux ont généré des traductions par une itération précédente : *adage* a généré les traductions françaises *adage* et *sentence* et *proverb* a généré les termes français *proverbe* et *sentence*. Le synset est donc instancié par ces traductions.

La proximité issue des espaces sémantiques indique alors que les termes *adage*, *proverbe* et *sentence* sont plus proches de *dicton* que de *scie*. Le candidat *dicton* est alors validé.

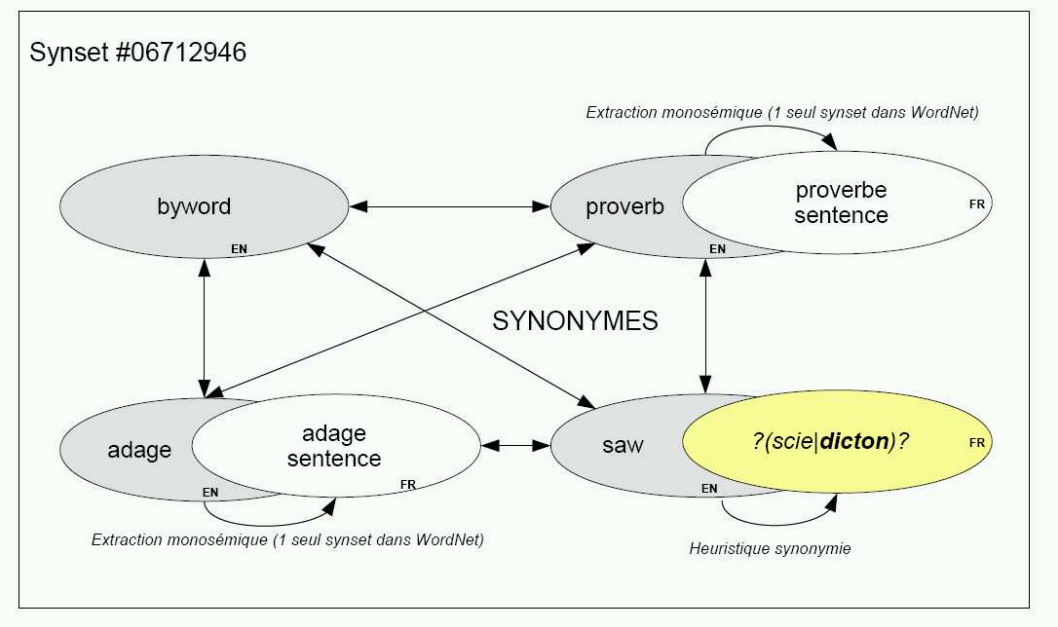



FIGURE 3.2 – Exemple de traduction exploitant l'heuristique fondée sur les relations de synonymie

**Hyponymie/Hyperonymie** On se propose également d'exploiter les relations d'hyponymie et d'hyperonymie pour déterminer quel est le candidat de traduction le plus adapté.

La figure 3.3 présente l'exemple du nom *fighter* en tant que *a high-speed military or naval airplane designed to destroy enemy aircraft in the air*[5]. Pour ce synset, l'humain peut intuitivement exploiter

4. *Une formule condensée mais inoubliable exprimant un fait important d'expérience qui est considéré comme vrai par beaucoup de gens*
5. *Un avion militaire ou naval à grande vitesse conçu pour détruire la force aérienne ennemie en vol*

la connaissances des traductions des différents synsets en relation d'hyperonymie et d'hyponymie pour déterminer que la meilleure traduction est *chasseur*. Nous proposons une caractérisation qui permettra au système de prendre cette décision en exploitant les mêmes connaissances.

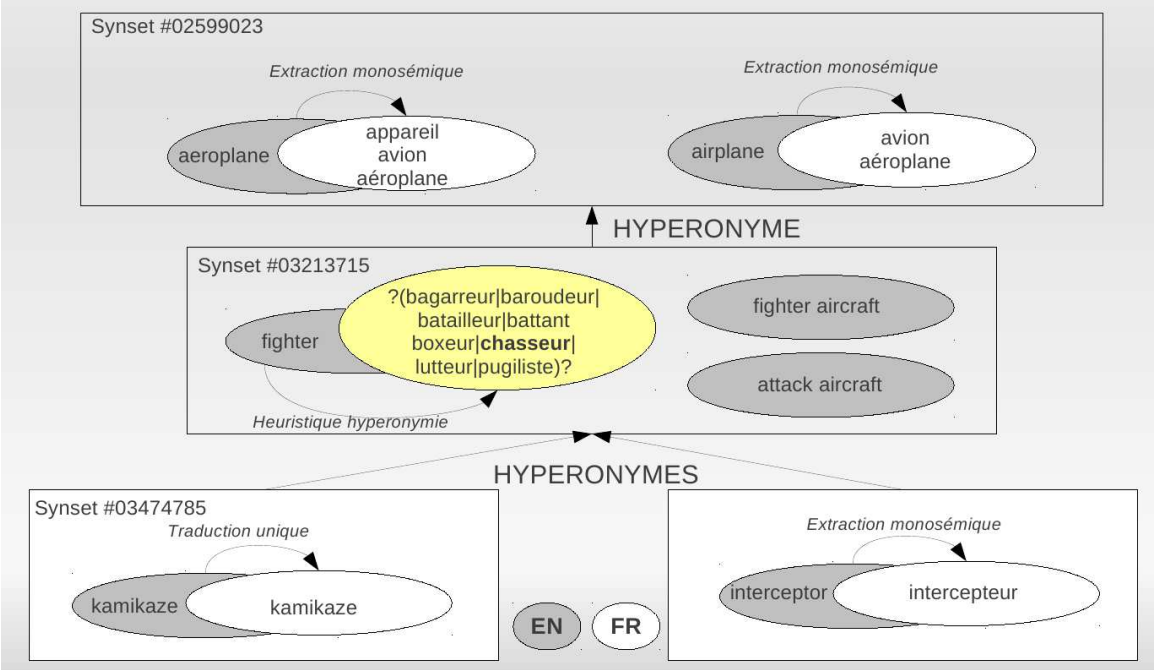



FIGURE 3.3 – Exemple de traduction exploitant l'heuristique fondée sur les relations d'hyperonymie

D'après [Vilnat 2005], la sémantique logique de [Sowa 1984] et la sémantique structurale issue des travaux de [Saussure (de) 1916] et fortement reprise par [Lyons 1970] trouvent une certaine similarité dans la définition de l'hyperonymie. En effet, pour les structuralistes, *le sémème hyponymique présente tous les sèmes qui sont également propres au sémème hyperonymique et au moins un sème qui va le spécifier et éventuellement le différencier des autres sémèmes de la catégorie.* [6] Ainsi, comme le montre la figure 3.4, l'ensemble des sèmes du mot le plus général est inclus dans l'ensemble de sèmes du mot le plus spécifique.

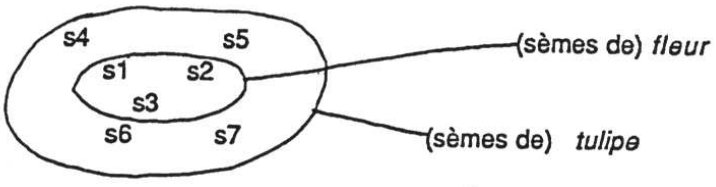



FIGURE 3.4 – Hyperonymie en analyse intentionnelle [Kleiber & Tamba 1990]

6. [Vilnat 2005], page 34

En sémantique logique, l'hyperonymie telle que [Sowa 1984] la définit repose sur une relation d'équivalence logique entre un ensemble intentionnel regroupant les caractéristiques des sémèmes et un ensemble extensionnel regroupant les entités référées. Cela permet de représenter un hyponyme par le graphe de définition de son hyperonyme (contenant les sémèmes propres à celui-ci), auquel on adjoint ou dont on modifie une ou plusieurs caractéristique(s).

Partant du postulat selon lequel un mot spécifique possède des caractéristiques plus complètes que son hyperonyme, nous émettons la double hypothèse suivante :

1. les contextes syntaxiques d'un mot général apparaissent souvent comme contexte syntaxique de ses hyponymes (on trouve par exemple *la vitesse du véhicule*, et *la vitesse du train, du bateau, du camion, de l'avion, du chasseur*)
2. l'éventail des contextes syntaxiques d'un mot spécifique est plus grand que ceux de ses hyperonymes (on trouve par exemple *la quille du bateau* mais pas *la quille du véhicule*, *la caténaire du train/tramway*, mais rarement la *caténaire du véhicule*, *le tir du chasseur*, mais rarement *le tir de l'avion*).

À partir de ces deux hypothèses, on cherche à définir un score nous donnant une estimation de la fiabilité de chaque candidat. Soit un synset $S_0$ possédant un ensemble non nul de synsets hyponymes instanciés $h(S_0)$ et un ensemble non nul de synsets hyperonymes instanciés $H(S_0)$, on calcule pour chaque terme candidat $c$ un score $\sigma(c)$ composé de deux termes.

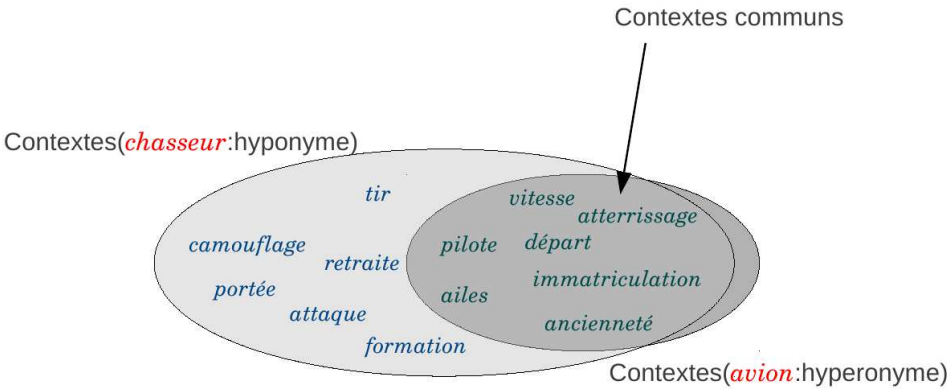



FIGURE 3.5 – Contextes d'un hyponyme et de son hyperonyme dans l'espace "Complémenté du nom"

Le premier terme correspond au score attribué par le(s) hyponyme(s) du synset à traduire. D'après l'hypothèse (1), on souhaite que les contextes syntaxiques du candidat $c$ qui sera choisi apparaissent

souvent comme contextes syntaxiques de ses hyponymes $h(S_0)$. Ceci est équivalent à dire que l'on souhaite que le plus grand nombre possible de contextes du candidat $c$ appartiennent à l'ensemble des contextes de chacun de ses hyponymes $h(S_0)$.

On voit sur la figure 3.5 que cela revient à faire en sorte que le nombre de contextes [7] communs au candidat et à chacun des termes hyponymes français soit le plus proche possible du nombre de contextes du candidat. Ceci nous mène à la définition de la première partie du score comme étant la moyenne de ces ratios pour tous les termes hyponymes français :

$$\sigma_{hypo}(c) = \frac{1}{|h(S_0)|} \sum_{s \in h(S_0)} \frac{1}{|termesFR(s)|} \sum_{t \in termesFR(s)} \frac{|ctx(t) \cap ctx(c)|}{|ctx(c)|} \tag{3.2}$$

avec $ctx(x)$ l'ensemble des termes cibles contextes de $x$ et $termesFR(s)$ l'ensemble des traductions françaises validées du synset $s$.

Le deuxième terme du score est le pendant du premier terme en considérant cette fois-ci les hyperonymes du synset à traduire. Le raisonnement est le même à la différence que cette fois le candidat $c$ est considéré comme le terme le plus spécifique et non plus comme le terme général. Le score doit donc être plus élevé lorsque, pour tous ses hyperonymes $H(S_0)$, le nombre de contextes communs au candidat $c$ et aux termes hyperonymes français est le plus proche possible du nombre de contextes de l'hyperonyme. Ceci est donné par la seconde partie du score définie ci-dessous :

$$\sigma_{hyper}(c) = \frac{1}{|H(S_0)|} \cdot \sum_{s \in H(S_0)} \frac{1}{|termesFR(s)|} \sum_{t \in termesFR(s)} \frac{|ctx(c) \cap ctx(t)|}{|ctx(t)|} \tag{3.3}$$

D'après les équations 3.2 et 3.3, on déduit la caractérisation suivante : le système doit choisir le candidat maximisant le score $\sigma(c)$ tel que :

$$\begin{aligned} \sigma(c) = & \frac{1}{|h(S_0)|} \sum_{s \in h(S_0)} \frac{1}{|termesFR(s)|} \sum_{t \in termesFR(s)} \frac{|ctx(t) \cap ctx(c)|}{|ctx(c)|} \\ & + \frac{1}{|H(S_0)|} \cdot \sum_{s \in H(S_0)} \frac{1}{|termesFR(s)|} \sum_{t \in termesFR(s)} \frac{|ctx(c) \cap ctx(t)|}{|ctx(t)|} \end{aligned}$$

L'hypothèse (2) sert à limiter respectivement les diviseurs à $|ctx(c)|$ et $|ctx(t)|$ et non $|ctx(c) \cup ctx(t)|$.

Nous utilisons cette heuristique distinctement sur les espaces sémantiques de complément du nom, sujet-verbe, et objet-verbe, en l'appelant respectivement *Hc*, *Hs* et *Ho*. En effet, ces espaces sont les espaces les plus riches dont nous disposons, car les trois relations syntaxiques sus-citées sont les trois relations les plus fréquentes en corpus parmi les relations entre un nom et un autre mot plein

7. on parle ici de la variété des contextes, on fait donc abstraction de leur nombre d'occurrences

(nom, verbe, adjectif, adverbe). Pour les *expressions multimots*†, le premier mot de l'expression est utilisé.

**Méronymie/Holonymie** Les relations de méronymie ou d'holonymie de type *Partie*[8] avec des synsets déjà instanciés peuvent également être exploitées pour déterminer le meilleur candidat cible. Notre hypothèse est qu'un concept faisant partie d'un autre est fortement susceptible d'apparaître dans ses cooccurrents par la relation complément du nom : *la pédale du vélo*, *le toit de l'immeuble*. Pour d'autre langue que le français, la relation peut-être différente (i.e. *bicycle pedal*). Cette heuristique est discutable car certaines prépositions formant la relation de complément du nom ne réalisent que rarement la relation de méronymie dans le sens proposé. De plus, même si la préposition *de* correspond parfois à notre caractérisation, ce n'est pas toujours le cas (*tour du monde, coup de vent...*).

L'ensemble des candidats étant restreint par les traductions des termes source, cette heuristique peut néanmoins permettre le choix du bon candidat. Le score d'un candidat est alors la moyenne des scores prenant en compte le nombre d'occurrences de la relation *complément du nom* entre chaque méronyme (ou holonyme) et le candidat, divisé par le nombre d'occurrences du candidat et du méronyme (ou holonyme) en position de complément du nom. Les candidats ayant les plus hauts scores sont conservés pour traduction. Cette heuristique sera notée $M$ dans les résultats.

La figure 3.6 présente l'exemple du nom *wash* dans le synset dont la définition est *the work of cleansing (usually with soap and water)*[9]. Ce terme est traduit par notre système grâce à l'heuristique exploitant les relations de méronymie/holonymie.

De plus, on remarque que la caractérisation utilisée pour la relation d'hyponymie est aussi applicable au cas de la méronymie (notée *Mh*). En effet beaucoup des méronymes présents dans PWN sont des méronymes possédant la même sémantique que leur holonyme mais de façon plus spécifique. Retenons pour exemple, le *crépuscule* méronyme du *soir*, la *courette* méronyme de la *maison* ou encore l'*inspiration* méronyme de la *respiration*. En outre, la figure de style synecdoque[10] est relativement souvent utilisée. Dans l'exemple : *Il a trouvé un toit pour passer la nuit.*, le *toit* peut être considéré comme un hyponyme de *logement* et possède un contexte plus varié. La caractérisation précédente est donc adaptée également à ces cas-là.

**Distance de Levenshtein** Nous appliquons une dernière heuristique (notée $L$). Si la langue cible est suffisamment proche de la langue source, la racine étymologique d'un mot peut être conservée d'une langue à l'autre, ainsi que différents sens issus de cette racine.

C'est le cas par exemple du mot anglais *unit* dont les sens de WordNet sont les suivants :

1. *unit of measurement, unit* ;

8. WordNet contient des relations de méronymie/holonymie de trois types : *Partie*, *Membre* ou *Substance*
9. *Le travail de nettoyage (généralement avec du savon et de l'eau)*
10. La partie pour le tout, La matière pour l'objet

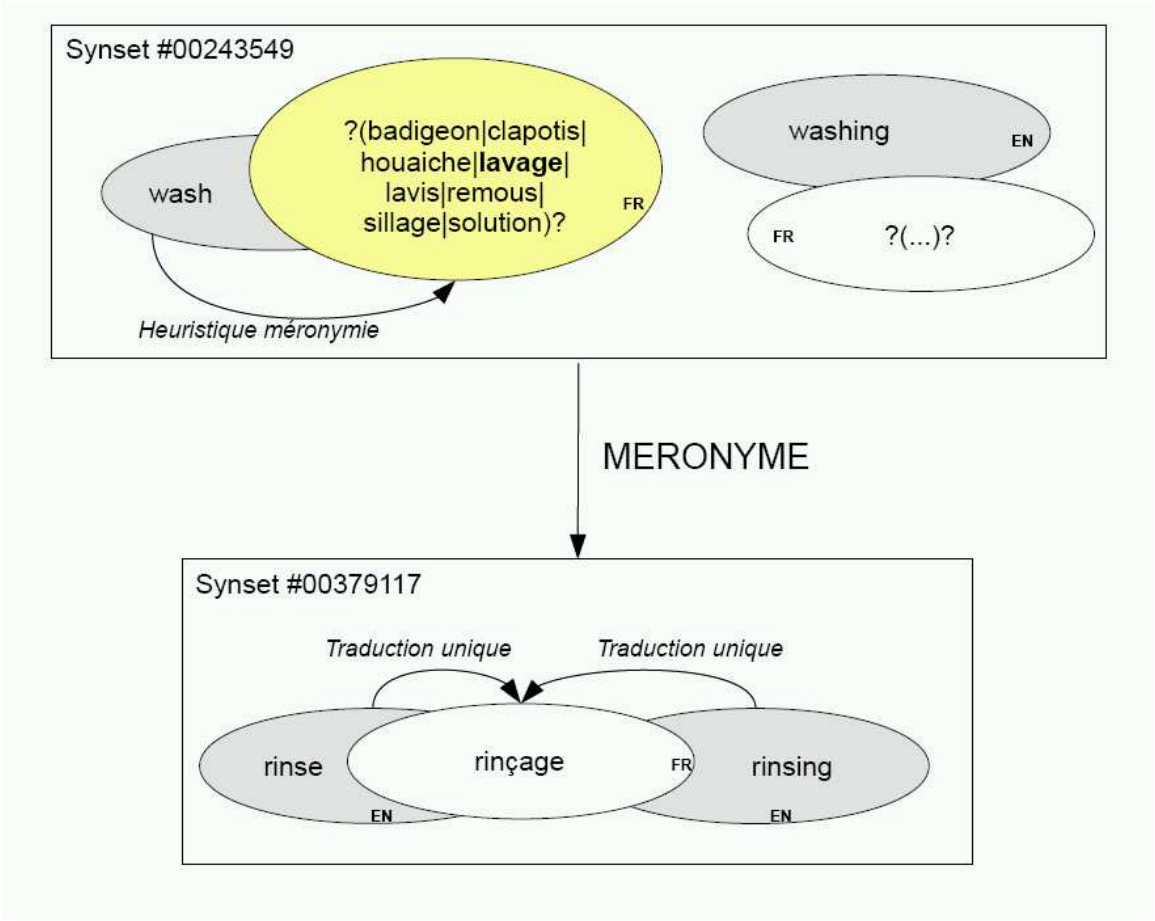



FIGURE 3.6 – Exemple de traduction exploitant l'heuristique fondée sur les relations d'holonymie

2. *unit (an individual or group or structure or other entity regarded as a structural or functional constituent of a whole)* ;
3. *unit, social unit* ;
4. *unit (a single undivided whole).*

Le mot français *unité* convient comme traduction pour chacune de ces définitions.

Cette heuristique consiste alors à valider le candidat dont la distance de Levenshtein avec le mot source est la plus faible, à condition que celle-ci soit inférieure à un certain seuil.

### 3.1.2 Évaluation

Afin de valider notre approche et nos hypothèses, nous nous comparons à la ressource Wolf qui présente l'intérêt d'avoir déjà été évaluée par rapport à EuroWordNet à la fois automatiquement et manuellement ([Sagot & Fišer 2008]). La version de JAWS évaluée est donc construite à partir de la version du PWN 2.0 utilisée par Wolf.

#### 3.1.2.1 Évolution des heuristiques

Nous définissons la *précision relative* comme étant la précision en considérant Wolf comme étant la vérité-terrain. Cette précision relative est donc le pourcentage de paires *(nom polysémique, synset)* correspondant à une entrée de Wolf lorsque ce synset existe dans Wolf, ce qui n'est pas toujours le cas.

Il y a deux raisons pour lesquelles nous préférons parler de précision relative et non d'une précision absolue. D'une part, la précision de Wolf n'est elle-même pas de 100% : elle est estimée à 77% sur les termes nominaux polysémiques par les auteurs. D'autre part, Wolf n'étant pas exhaustif sur l'ensemble du vocabulaire français, nous souhaitons analyser manuellement l'ensemble des paires trouvées par notre système n'apparaissant pas dans Wolf, à la fois lorsque le synset est présent dans Wolf et lorsqu'il ne l'est pas.

Le procédé itératif employé consiste à évaluer pour chaque heuristique le sous-ensemble des paires produites qui n'existaient pas à l'itération précédente et à conserver le sous-ensemble donnant la meilleure précision relative. Nous parlerons de *parties additionnelles* pour faire référence à de tels sous-ensembles.

Le tableau 3.3 montre les heuristiques qui sont choisies à chaque itération ainsi que la précision relative des parties additionnelles correspondant aux heuristiques choisies.

On remarque que ces différentes heuristiques s'enrichissent l'une l'autre. On observe par exemple à l'itération 4 le choix de l'heuristique $Mh$ donnant une précision relative de sa partie additionnelle de 30,8%, ce qui est supérieur à la meilleure partie additionnelle obtenue à l'itération précédente. Ceci est dû au fait qu'à l'itération précédente le système a généré un certain nombre de validations

| Itérations | Meilleure heuristique | Précision relative additionnelle |
|---|---|---|
| 0 | E | 0,305 |
| 1 | L | 0,455 |
| 2 | M | 0,354 |
| 3 | M | 0,300 |
| **4** | **Mh** | **0,308** |
| 5 | Hc | 0,281 |
| **6** | **Mh** | **0,300** |
| 7 | Hc | 0,246 |
| **8** | **Hs** | **1,000** |
| 9 | S | ,210 |

TABLE 3.3 – Heuristiques choisies et précisions relatives des parties additionnelles

de traduction lui permettant d'être plus pertinente lors de cette nouvelle itération. Ce phénomène se produit également aux itérations 6 et 8.

Les itérations s'arrêtent lorsque le nombre de paires présentes dans Jaws et Wolf à la fois n'augmente plus.

La courbe de la figure 3.7 permet de vérifier que le pourcentage de paires n'appartenant pas à Wolf et restant donc à évaluer manuellement se maintient suffisamment proche du pourcentage obtenu après la première itération.

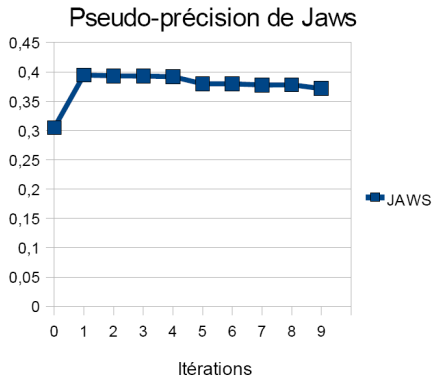



FIGURE 3.7 – Évolution de la précision relative de l'ensemble de la ressource en cours de construction

On peut également suivre à la figure 3.8 l'évolution du nombre de paires (nom polysémique, synset) de la ressource conservée. On observe sur cette figure le nombre final de paires (*nom polysémique, synset*) atteint. Celui-ci attteint plus de la moitié du nombre de paires présentes dans le WordNet source et plus du double du nombre de paires présentes dans Jaws. Il reste maintenant à évaluer la qualité des paires ne figurant pas dans Wolf.

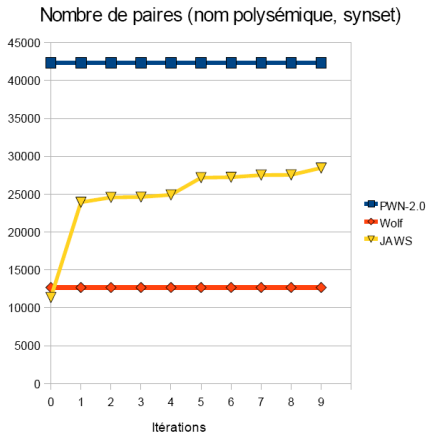



FIGURE 3.8 – Évolution du nombre de paires (nom polysémique, synset) au fur et à mesure des itérations

#### 3.1.2.2 Évaluation de la meilleure séquence

Nos résultats montrent que l'extraction pure produit moins de traductions que ce que l'on trouve dans Wolf (26,9 % contre 30,0 % des synsets de PWN). En revanche, dès la première itération, on obtient un plus grand nombre de traductions que Wolf.

Nous mesurons d'une part la couverture obtenue par Wolf et JAWS en pourcentage du nombre de synsets polysémiques de PWN.

D'autre part, nous classons les paires *Terme-Synset* obtenues dans la ressource cible en trois catégories. Dans la catégorie 1, le synset concerné est présent dans Wolf, et le terme français également.

Dans la catégorie 2, P le synset concerné est présent dans Wolf mais les termes de Wolf ne contiennent pas le terme proposé par JAWS. C'est le cas par exemple pour les termes français du synset contenant les termes anglais *shade, tincture, tint, tone* : JAWS propose les termes *coloris, couleur, lasure, nuance, teinte, teinture, ton* tandis que Wolf propose les termes *ordinaire, teinte, variables*, ou encore pour le synset anglais *hospital, infirmary* pour lequel JAWS propose les termes *hosto, hôpital, infirmerie* tandis que Wolf propose uniquement *hospital*. Ces exemples sont choisis volontairement pour montrer que l'analyse manuelle est nécessaire. En effet, l'absence d'une paire *terme-synset* dans Wolf ne signifie pas nécessairement qu'il s'agit d'une erreur.

Enfin, dans la catégorie 3, le synset concerné n'a pas de traduction dans Wolf. L'évaluation manuelle est donc également nécessaire.

Les résultats sont présentés dans le tableau 3.4. La meilleure séquence itérative produit 67,3 % du nombre de synsets polysémiques de PWN avec 12,5 % des paires présentes dans Wolf (précision des

termes nominaux polysémiques de Wolf estimée à 77 % par leurs auteurs).

| | *Paires traduites* | Cat1. $P \in Wolf$ | Cat2. $P \notin Wolf$ | Cat3. $S \notin Wolf$ |
|---|---|---|---|---|
| Wolf | 30 % | | | |
| Extraction | 26,9% | 8,3%(30,5%) | 18,8%(69,6%) | 73,0% |
| Meilleure séquence | 67,3% | 12,5%(37,2%) | 21,1%(62,9%) | 66,5% |

TABLE 3.4 – Pourcentage des paires nominales polysémiques traduites et répartition des paires sur 3 catégories. Entre parenthèses figure le cas où l'on considère uniquement les synsets appartenant à Wolf.

Wolf ne répertoriant pas exhaustivement l'ensemble des paires possibles pour un synset, nous procédons à l'analyse manuelle d'un extrait aléatoire des paires de catégorie 2 et 3.

#### 3.1.2.3 Catégorie 2 : la paire n'appartient pas à Wolf

Nous proposons de classer les différences entre Wolf et JAWS selon le tableau 3.5.

| | | | |
|---|---|---|---|
| **Cat.2** | Manque Partiel (MP) dans Wolf : au moins une traduction de S mais pas de traduction de T | MP1 | Traduction JAWS correcte |
| | | MP2 | Traduction JAWS incorrecte |
| | Différence (D) de traduction | D1 | La traduction de Wolf est incorrecte et celle de JAWS est correcte |
| | | D2 | La traduction de Wolf est moins bonne |
| | | D3 | Les deux traductions sont correctes et équivalentes |
| | | D4 | La traduction de JAWS est moins bonne |
| | | D5 | La traduction de JAWS est incorrecte et celle de Wolf est correcte |
| | Non résolu (*Wrong - W*) | W | Aucune traduction n'est adaptée |
| **Cat.3** | Manque Total (MT) dans Wolf : aucune traduction de S | MT1 | Traduction JAWS correcte |
| | | MT2 | Traduction JAWS incorrecte |

TABLE 3.5 – Différences avec Wolf pour une paire P constituée d'un synset S et d'un terme T

Le tableau 3.6 montre l'analyse manuelle (sur un échantillon de 40 paires) des paires absentes de Wolf mais présentes dans JAWS pour les synsets présents dans Wolf (21,1 % des paires concernant ces synsets). Afin d'estimer la précision de cette catégorie, nous considérons les cas où nous avons trouvé ces paires meilleures ou équivalentes à celles de Jaws, c'est à dire en suivant notre classification

des différences, les cas $(MP1, D1, D2, D3)$. Cette estimation de la précision des paires de catégorie 2 est donnée en dernière colonne du tableau. On constate que les valeurs sont assez semblables pour la ressource issue de l'extraction seule et pour la ressource issue de la meilleure séquence : 67,5 % de ces paires sont considérées meilleures ou équivalentes à celles de Wolf.

| | MP1 | MP2 | D1 | D2 | D3 | D4 | D5 | W | MP1+D1+D2+D3 |
|---|---|---|---|---|---|---|---|---|---|
| **Extraction** | 20 | 5 | 3 | 0 | 4 | 1 | 6 | 1 | $67,5\ \% \pm 14,5$ |
| **Meilleure séquence** | 25 | 10 | 1 | 0 | 1 | 2 | 1 | 0 | $67,5\ \% \pm 14,5$ |

TABLE 3.6 – Analyse des paires de catégorie 2 ($P \notin Wolf$) sur un échantillon de 40 paires

#### 3.1.2.4 Catégorie 3 : le synset n'existe pas dans Wolf

Le tableau 3.7) présente les résultats de l'évaluation manuelle sur les paires de catégorie 3, c'est-à-dire lorsque les synsets sont totalement absents de Wolf. Leur analyse manuelle sur un nouvel échantillon de 40 paires montre qu'elles sont correctes à 65,0 %.

| | MT1 | MT2 |
|---|---|---|
| **Extraction** | $82,5\ \% \pm 11,8$ | $17,5\ \% \pm 11,8$ |
| **Meilleure séquence** | $65,0\ \% \pm 14,8$ | $35,0\ \% \pm 14,8$ |

TABLE 3.7 – Analyse des paires de catégorie 3 ($S \notin Wolf$)

Enfin le tableau 3.8 donne la micro-précision estimée sur l'ensemble des catégories à l'aide de Wolf et des validations manuelles. Nous connaissons le pourcentage de paires appartenant à chaque catégorie ainsi que la précison estimée sur chacune d'entre elles. Cela nous permet de faire une estimation globale sur l'ensemble de la ressource par l'intermédiaire de la formule suivante :

$$\sum_{i \in \{1,2,3\}} Précision(Cat(i)) * Pourcentage(paire \in Cat(i))$$

D'après les résultats du tableau,nous avons donc pour la meilleure séquence :

$$0.125 * 77 + 0.211 * 67,5 + 0.671 * 65 = 67,1$$

#### 3.1.2.5 Analyse des résultats

Après 9 itérations des heuristiques, nous obtenons la meilleure ressource avec une couverture de 67,3 % du nombre de paires d'origine. La ressource obtenue contient un total de 28 464 termes nominaux uniques, et ceci avec une précision estimée à 67 % pour les termes nominaux polysémiques. On

| | **Extraction** | | **Meilleure séquence** | |
|---|---|---|---|---|
| | %age | Précison | %age | Précision |
| **Catégorie 1** | 8,3 | 77,0 | 12,5 | 77,0 |
| **Catégorie 2** | 18,8 | 67,5 $\pm$14,5 | 21,1 | 67,5 $\pm$14,5 |
| **Catégorie 3** | 73,0 | 82,5 $\pm$11,8 | 66,5 | 65,0 $\pm$14,8 |
| **Ressource totale** | 100,0 | 79,3 $\pm$11,3 | 100,0 | 67,1 $\pm$12,9 |

TABLE 3.8 – Estimation de la précision sur l'ensemble de la ressource

obtient au final un WordNet français couvrant plus deux fois plus de synsets nominaux polysémiques que Wolf pour une perte de précision estimée à 10 points.

L'heuristique fournissant les meilleurs résultats à la première itération est celle exploitant la distance de Levenshtein. Ceci peut s'expliquer par le fait qu'un faible nombre de synsets sont instanciés avant la première itération, ceci rendant difficile l'exploitation des autres heuristiques.

Un des inconvénients de la méthode proposée réside dans l'incapacité du système à ne choisir aucun candidat parmi les traductions proposées.

Si le dictionnaire bilingue fournit un certain nombre de candidats mais ne fournit pas de traduction pour un des sens WordNet du terme source, la traduction de celui-ci sera nécessairement fausse. C'est notamment le cas pour les mots ne possédant pas d'équivalent français, mais nous avons aussi des exemples où les dictionnaires bilingues utilisés font défaut. Nous avons par exemple le mot anglais *flagship* dont les deux sens que WordNet répertorie sont les suivants :

1. *the chief one of a related group; "it is their flagship newspaper"* ;
2. *the ship that carries the commander of a fleet and flies his flag.*

*Flagship* n'a pour seul équivalent français dans nos dictionnaires que le mot *fleuron* correspondant à la définition du sens 1 mais pas à celle du sens 2 qui pourrait plutôt se traduire par *navire amiral*[11].

Si le candidat le plus correct ne figure pas dans les entrées de l'espace sémantique (comme les noms propres dans notre cas), la traduction choisie sera nécessairement fausse.

Nous ne pouvons pas quantifier clairement la proportion de ces cas d'erreurs mais nous pensons que ces deux raisons expliquent un bon nombre des cas d'erreur obtenus. La méthode gagnerait donc à fixer quelques critères de non-choix de candidat, éventuellement par le calcul d'un score de confiance propre à chaque heuristique.

### 3.1.3 Perspectives et conclusions

Dans un premier temps, nous analysons de façon quantitative les relations sémantiques présentes dans WordNet dans le but de déterminer si certaines informations qui y sont présentes font défaut

11. `http://fr.wikipedia.org/wiki/Navire_amiral`

à notre traitement. Dans un deuxième temps, nous évoquons plus largement les différentes pistes que nous pensons susceptibles d'améliorer notre système de traduction de WordNet. Enfin, nous concluons cette section par un récapitulatif des ressources produites.

#### 3.1.3.1 Analyse quantitative des relations sémantiques présentes dans WordNet

Nous avons mené une analyse quantitative des relations sémantiques présentes dans WordNet. La plupart des travaux utilisant WordNet exploite les relations de synonymie et d'hyperonymie, mais beaucoup d'autres relations y sont présentes et restent plus rarement exploitées.

Pour chaque catégorie syntaxique et chaque relation sémantique, nous avons rapporté dans les figures 3.9, 3.10, 3.11 et 3.12, le nombre moyen de relations sémantiques gouvernant chaque sens de chaque mot (i.e. un mot dans un synset donné), ainsi que chaque sens de chaque mot polysémique.

Par exemple, nous voyons à la première ligne du tableau 3.9 qu'il y a en moyenne 1,39 synonymes par sens et par mot dans WordNet-2.0 et 1,41 dans WordNet-3.0 (on compte les synonymes de *mistake* dans chacun de ses 3 synsets, et on compte également les synonymes de *error* dans chacun de ses 7 synsets).

Certains types de relations se déclinent en plusieurs variantes comme c'est le cas pour les relations méronymes/holonymes, qui peuvent être de type *substance, membre* ou *partie*. Les exemples de la dernière colonne aident à élucider la nature des relations qui pourraient être inconnues.

Enfin, les deux dernières lignes de chaque tableau donnent d'une part la moyenne du nombre de relations par sens et par mot, toutes relations confondues, et d'autre part, le nombre moyen de ces relations si on ne considère que les termes polysémiques. En effet, puisque ce sont ces termes particulièrement que nous essayons de traduire, il est intéressant de voir les statistiques qui les concernent.

Dans notre étude du système de traduction, nous avons exploité la relation de synonymie, la relation d'hyponymie/hyperonymie, et, fait suffisamment rare pour être remarqué la relation de méronymie de partie. Cela représente 75% du nombre moyen de relations par nom. La seule relation non exploitée ayant une proportion non négligeable est la relation de dérivation avec les verbes, ce qui représenterait 7% de plus. Des futurs travaux dans cette direction sont donc fortement encouragés.

Nous observons qu'il existe pour les noms d'autres relations sémantiques que nous n'avons pas exploitées : principalement la méronymie/holonymie de type membre ou substance, l'antonymie, et les dérivés étymologiques, notamment ceux de type verbaux... Ces tableaux montrent par ailleurs une plus forte densité de relations sémantiques pour les verbes : 8,87 pour les verbes polysémiques dans WordNet-3.0 comparé à 1,73 pour les noms polysémiques dans WordNet-3.0. Cela nous donne de bons espoirs pour un procédé de traduction similaire à celui utilisé ici pour les noms. Il serait

| Relations | | WN-2.0 | WN-3.0 | Exemple |
|---|---|---|---|---|
| **Synonymes** | | 1.39 | 1.41 | *mistake* ↔ *error* |
| **Antonymes** | | 0.03 | 0.03 | *bravery* ↔ *cowardice* |
| **Hyperonymes** | Subsomption | 1.03 | 0.91 | *harmony* → *sound property* |
| | Instance | 0 | 0.12 | *Isle of Wight* → *isle* |
| **Hyponymes** | Subsomption | 1.12 | 1.02 | *sound property* → *harmony* |
| | Instance | 0 | 0.11 | *isle* → *Isle of Wight* |
| **Holonymes** | Partie | 0.10 | 0.11 | *Andalucia* → *Spain*<br>*heliosphere* → *Milky Way*<br>*ozone hole* → *ozone_layer* |
| | Membre | 0.18 | 0.18 | *planet* → *solar system*<br>*ally* → *coalition* |
| | Substance | 0.10 | 0.11 | *carbon* → *coal*<br>*cannabis* → *joint*<br>*amino acid* → *protein* |
| **Méronymes** | Partie | 0.14 | 0.15 | *Pakistan* → *Hindu Kush Mountains*<br>*Pakistan* → *Islamabad*<br>*moon light* → *moon ray* |
| | Membre | 0.20 | 0.20 | *Pakistan* → *Pakistani*<br>*battalion* → *company* |
| | Substance | 0.14 | 0.15 | *natural gas* → *methane*<br>*bread* → *flour*<br>*stalagmite* → *dripstone* |
| **Attributs** | | 0.01 | 0.01 | *financial condition* → *poor*<br>*financial condition* → *rich*<br>*fear* → *afraid* |
| **Termes du domaine** | Noms | 0.06 | 0.07 | *electricity* → *earth*[1]<br>*electricity* → *amplification*[2] |
| | Verbes | 0.02 | 0.02 | *telephone* → *call in*<br>*telephone* → *hold on* |
| | Adjectifs | 0.02 | 0.01 | *electricity* → *voltaic*<br>*electricity* → *polyphase* |
| | Adverbes | 0.00 | 0.00 | *music* → *allegro*<br>*sport* → *at home*<br>*sailing* → *close to the wind* |
| **Termes d'usage** | Noms | 0.01 | 0.02 | *trademark* → *Alka − seltzer*<br>*acronym* → *scuba*[3]<br>*plural form* → *telecommunication*<br>*figure of speech* → *rainy day*[4] |
| | Verbes | 0.00 | 0.00 | *portmanteau words* → *dandle* |
| | Adjectifs | 0.00 | 0.00 | *slang* → *uncool*<br>*euphemism* → *gone*[5] |
| | Adverbes | 0.00 | 0.00 | *colloquial* → *okay*<br>*superlative* → *nearest* |
| **Termes régionaux** | **All** | 0.04 | 0.04 | *Australia* → *swagman*[6]<br>*United Kingdom* → *scrimshank*[7]<br>*India* → *castless*<br>*Scotland* → *langsyne*[8] |
| **...** | | ... | ... | ... |

[1] Définition : *a connection between an electrical device and a large conducting body, such as the earth (which is taken to be at zero voltage)*
[2] Définition : *(electronics) the act of increasing voltage or power or current*
[3] Définition : *a device (trade name Aqua-Lung) that lets divers breathe under water; scuba is an acronym for self-contained underwater breathing apparatus*
[4] Définition : *a (future) time of financial need; "I am saving for a rainy day"*
[5] Définition : *dead; "he is deceased"; "our dear departed friend"*
[6] Définition : *an itinerant Australian laborer who carries his personal belongings in a bundle as he travels around in search of work*
[7] Définition : *British military language: avoid work*
[8] Définition : *at a distant time in the past (chiefly Scottish)*

| Relations | | WN-2.0 | WN-3.0 | Exemple |
|---|---|---|---|---|
| ... | | ... | ... | ... |
| **Domaine** | | 0.04 | 0.05 | *Quantum theory* → *physics* |
| **Registre de langue** | | 0.01 | 0.02 | *telecommunication* → *plural form*<br>*rainy day* → *figure of speech* |
| **Région** | | 0.01 | 0.01 | *old man*[9]→ *USA* |
| **Dérivés étymologiques** | Noms | 0 | 0.04 | *agent* → *agency* |
| | Verbes | 0.37 | 0.36 | *deity* → *deify* |
| | Adjectifs | 0 | 0.21 | *divinity* → *divine* |
| **Moyenne par sens de nom** | | 5.03 | 5.34 | |
| **Moyenne par sens de nom polysémique** | | 1.35 | 1.73 | |

[9] Définition : *boss*

TABLE 3.9 – WordNet : nombre moyen de relations sémantiques par nom et par synset

| Relations | | WN-2.0 | WN-3.0 | Exemple |
|---|---|---|---|---|
| **Synonymes** | | 1.72 | 1.72 | *expire* ↔ *exhale* |
| **Antonymes** | | 0.11 | 0.11 | *expire* ↔ *inspire* |
| **Hyperonymes** | | 0.95 | 0.95 | *caramelize* → *convert* |
| **Hyponymes** | | 1.45 | 1.46 | *convert* → *caramelize* |
| **Voir aussi** | | 0 | 0 | *falsify* → *make up* |
| **Entailment** | | 0.04 | 0.04 | *dream* → *sleep* |
| **Cause** | | 0.02 | 0.02 | *trigger* → *happen* |
| **Groupe de verbes** | | 0.16 | 0.15 | *summarize*[1]<br>→ *summarize*[2] |
| **Domaine** | | 0.07 | 0.07 | *freeze*[3]→ *surgery* |
| **Registre de langue** | | 0.00 | 0.00 | *keep one's eyes open*<br>→ *colloquial* |
| **Région** | | 0.00 | 0.00 | *scrimshank*[4]<br>→ *United Kingdom* |
| **Dérivés étymologiques** | Noms | 2.27 | 2.25 | *observe* → *observation* |
| | Adjectifs | 0 | 0.18 | *observe* → *observation* |
| **Moyenne par sens de verbe** | | 6.78 | 6.94 | |
| **Moyenne par sens de verbe polysémique** | | 8.64 | 8.87 | |

[1] Définition : *be a summary of*
[2] Définition : *give a summary of*
[3] Définition : *anesthetize by cold*
[4] Définition : *British military language: avoid work*

TABLE 3.10 – WordNet : nombre moyen de relations sémantiques par verbe et par synset

| Relations | | WN-2.0 | WN-3.0 | Exemple |
|---|---|---|---|---|
| **Synonymes** | | 1.40 | 1.37 | *incorrect* ↔ *wrong* |
| **Antonymes** | | 0.17 | 0.18 | *relative* ↔ *absolute* |
| **Attribut d'un nom** | | 0.03 | 0.03 | *afraid* → *fear* |
| **Participe d'un verbe** | | 0.01 | 0.00 | *seeking* → *seek*<br>*played* → *play* |
| **Pertainyme** | | 0.24 | 0.25 | *rupestral* → *rock, stone* |
| **Similaire à** | | 1.13 | 1.12 | *agitated* ↔ *excited* |
| **Voir aussi** | | 0 | 0 | *ambiguous* ↔ *unclear* |
| **Domaine** | | 0.05 | 0.05 | *compatible*[1] → *computer* |
| **Registre de langue** | | 0.01 | 0.01 | *earlier* → *comparative*<br>*freaky* → *slang*<br>*fin de siecle* → *French* |
| **Région** | | 0.00 | 0.00 | *castless* → *India* |
| **Dérivés étymologiques** | Noms | 0 | 0.98 | *exclusive* → *exclusiveness* |
| | Verbes | 0 | 0.12 | *exclusive* → *exclude* |
| | Adverbes | 0 | 0.00 | *feasable* → *feasably* |
| **Moyenne par sens d'adjectif** | | 3.04 | 4.12 | |
| **Moyenne par sens d'adjectif polysémique** | | 1.35 | 2.01 | |

[1] Définition : *capable of being used with or connected to other devices or components without modification*

TABLE 3.11 – WordNet : nombre moyen de relations sémantiques par adjectif et par synset

| Relations | | WN-2.0 | WN-3.0 | Exemple |
|---|---|---|---|---|
| **Synonymes** | | 1.18 | 1.14 | *badly* ↔ *poorly* |
| **Antonymes** | | 0.22 | 0.22 | *badly* ↔ *well* |
| **Relation adjectif** | | 1.31 | 1.33 | *contextually* → *contextual* |
| **Domaine** | | 0.01 | 0.01 | *pizzicato* → *music* |
| **Registre de langue** | | 0.03 | 0.03 | *roughly* → *colloquial* |
| **Région** | | 0.00 | 0.00 | *langsyne*[1] → *Scotland* |
| **Dérivés étymologiques** | Adjectifs | 0 | 0.00 | *feasably* → *feasable* |
| **Moyenne par sens d'adverbe** | | 2.75 | 2.72 | |
| **Moyenne par sens d'adverbe polysémique** | | 0.46 | 0.46 | |

[1] Définition : *at a distant time in the past (chiefly Scottish)*

TABLE 3.12 – WordNet : nombre moyen de relations sémantiques par adverbe et par synset

certainement intéressant de traduire simultanément les noms et les verbes afin d'exploiter également les relations intercatégorielles telle que *dérivés étymologiques*. Dans cet exemple additionnel de relation à exploiter, une heuristique distributionnelle pourrait assez facilement être proposée. Il s'agirait par exemple de choisir dans un espace sémantique de type fenêtre (mixant ainsi les catégories grammaticales) le verbe le plus proche du nom en relation (ou vice-versa).

#### 3.1.3.2 Pour aller plus loin

Un certain nombre d'idées pour l'amélioration de cette traduction automatique nous sont apparues au cours de l'évaluation et de l'analyse des résultats. Nous mentionnons ici celles que nous percevons comme les plus pertinentes et qu'il faudrait rapidement mettre en œuvre.

Bien que toutes les heuristiques ne soient pas utilisées dans la séquence optimale, on remarque qu'elles produisent chacune des résultats intéressants. Nous proposons donc d'étudier dans des travaux ultérieurs le gain éventuel en précision apporté par une utilisation combinée (et non séquentielle) des différentes heuristiques. Il pourrait s'agir dans un premier temps de conserver tous les lemmes cibles produits par au moins deux heuristiques différentes et de réitérer.

Par ailleurs, nous remarquons par l'analyse des relations sémantiques présentes dans WordNet que les verbes polysémiques sont liés au réseau lexical par beaucoup plus de relations sémantiques que ne le sont les noms de WordNet. En effet, dans WordNet-2.0, les verbes polysémiques possèdent en moyenne 6,78 relations sémantiques, contre 5,03 pour les noms polysémiques. Si on se restreint aux termes polysémiques, ce nombre devient plus faible pour les noms (1,35). En revanche, il est encore plus élevé pour les verbes puisqu'il atteint une moyenne de 8,64. Ceci s'explique par le fait que la majorité des verbes de WordNet sont polysémique (65%) tandis que seulement 49% des noms ne le sont.

Il serait certainement judicieux de traduire les verbes polysémiques car ils possèdent beaucoup plus de relations sémantique que les noms polysémiques dans WordNet. On peut ainsi espérer qu'en utilisant des heuristiques caractérisant les relations majoritaires des verbes et en exploitant ainsi plus d'information sémantique que ça n'a été le cas pour les noms, la précision des traductions obtenues sera assez élevée. Si celle-ci s'avère être plus élevée que celle des noms, on pourrait également utiliser la relation de dérivation liant les noms aux verbes afin de mieux traduire les noms (ou vice-versa).

Les adjectifs polysémiques ne sont pas en reste, possédant également un nombre plus important de relations que les noms polysémiques (2,01 contre 1,73 dans WordNet-3.0 ou un nombre égal dans WordNet-2.0). Leur traduction par une méthode de ce type n'est donc pas exclue. Enfin, la plupart des relations des adverbes étant une relation directe de dérivation des adjectifs, la traduction de ceux-ci peut également être mise en place mais il restera un nombre important de cas non traitables par cette méthodologie.

Ce travail nous a également inspiré d'autres pistes plus exploratoires pour les tâches d'acquisition. Nous sommes consciente que ces heuristiques ne sont pas suffisamment robustes pour être utilisées directement dans les espaces sémantiques pour acquérir les relations sémantiques ainsi caractérisées (synonymie, hyperonymie/hyponymie, holonymie/méronymie). C'est ici la restriction des mots à la liste des candidats de traduction qui permet de contraindre l'espace des candidats et de déterminer quel mot est synonyme, hyperonyme ou méronyme d'un autre.

En revanche, il serait intéressant d'analyser de façon plus systématique les distributions syntaxiques caractérisant les relations sémantiques en utilisant le WordNet anglais de Princeton et des espaces sémantiques construits sur la langue anglaise. S'il existe réellement de telles caractérisations, cette analyse mènera à des caractérisations distributionnelles plus fines et plus exploitables pour l'acquisition de relations sémantiques dans les espaces distributionnels.

#### 3.1.3.3 Bilan

Le WordNet français ainsi obtenu couvre plus deux fois plus de paires (terme nominal polysémique, synset) que Wolf, avec une perte de précision estimée à 10 points sur les termes polysémiques. L'idéal serait maintenant de pouvoir combiner ces ressources par union (ou intersection) afin d'obtenir une ressource plus exhaustive (ou respectivement plus précise), en fonction de la tâche ou de l'application visée.

La méthode en elle-même peut être généralisée à d'autres langues, à condition que l'on dispose d'un dictionnaire bilingue riche, d'un analyseur syntaxique, et que la langue cible partage beaucoup de cognats avec la langue source (l'heuristique la plus efficace étant la distance de Levenshtein). Enfin, quelques modifications peuvent être nécessaires pour des langues dans lesquelles la structure de complément du nom ne s'emploierait pas de la même manière.

## 3.2 Induction automatique de sens

Nous pensons que la projection des mots dans des espaces vectoriels permet non seulement de retrouver une certaine similarité sémantique mais aussi de regrouper des ensembles de mots qui pourront définir les différents sens des mots polysémiques, à l'instar de [Schütze 1998], [Lin 1998], [Pantel & Lin 2002], [Ferret 2004] ou [Véronis 2004] dont nous avons décrit les travaux à la section 2.2.2.4. C'est l'induction de ces sens (*Word Sense Induction, WSI*) qui fait l'objet de la partie notre travail que nous abordons maintenant.

Notre travail se distingue des précédents par le type d'éléments que nous cherchons à clusteriser. Tandis que certains regroupent des cooccurrents ([Ferret 2004], [Véronis 2004]), d'autres des instances de contextes ([Schütze 1998]), et d'autres tous les mots du vocabulaire ([Lin 1998], [Pantel & Lin 2002]), nous cherchons à regrouper des plus proches voisins. Si le clustering de plus proches

voisins s'avère pertinent pour la distinction en sens, alors nous pourrons utiliser les éléments de nos clusters comme données d'apprentissage d'un algorithme de *classification k-nearest neighbours*† pour la tâche de désambiguïsation. En effet, après une analyse syntaxique identique à celle utilisée dans la constitution de nos espaces, chaque nouvelle instance ambiguë peut être représentée dans ceux-ci, mettant ainsi à disposition une distance entre les instances ambiguës et les éléments de nos clusters.

Notre travail se différencie également par la distinction de plusieurs espaces sémantiques dont nous combinons les spécificités. En effet, jusqu'à présent les travaux faisant usage des espaces sémantiques fondés sur les cooccurrences syntaxiques traitaient tous les contextes syntaxiques de la même façon, indépendemment du type de relation syntaxique formant le contexte (la seule distinction résidait dans le fait d'utiliser une relation ou non). Notre travail montre la variabilité des distances entre les mots selon les relations syntaxiques utilisées et explore les apports de la prise en compte spécifique des différentes relations.

Notre travail reprend en outre un algorithme de recherche rapide de plus proches voisins approximatifs développé par [Ravichandran *et al.* 2005] et le modifie pour d'une part réduire sa complexité algorithmique et d'autre part pour distribuer les calculs sur les noeuds d'un cluster.

Nous présentons tout d'abord les modifications apportées à l'algorithme de recherche rapide des plus proches voisins de [Ravichandran *et al.* 2005]. Nous montrons ensuite la variabilité des distances relatives des mots en fonction des différents espaces. Nous poursuivons par la description des méthodes de clustering adaptées pour prendre en compte les disparités de chaque espace. Enfin, nous terminons cette section par une mise en perspective de nos résultats et proposons quelques pistes d'exploration supplémentaires.

### 3.2.1 Recherche rapide des plus proches voisins approximatifs

Dans ses travaux sur les fonctions LSH sensibles à la similarité cosinus, [Charikar 2002] décrit également un algorithme de recherche rapide du plus proche voisin approximatif à partir des vecteurs de bits issus du hachage LSH. [Ravichandran *et al.* 2005] ont modifié cet algorithme de sorte qu'il soit capable de rechercher les $n$ plus proches voisins approximatifs. Cet algorithme est présenté dans notre état de l'art à la section 2.1.2.3.

Nous rappelons ici le fonctionnement de l'algorithme original. Nous introduisons également deux modifications que nous proposons pour réduire la complexité de l'algorithme pour l'utilisation spécifique que nous en faisons. Enfin nous décrivons brièvement l'implémentation parallèle que nous avons effectuée en MPI[12] afin de pouvoir utiliser l'algorithme sur un cluster de machines.

12. Message Passing Interface - http://www-unix.mcs.anl.gov/mpi/

#### 3.2.1.1 Méthode originale

La méthode de [Charikar 2002] consiste à tirer aléatoirement $p$ permutations de $d$ éléments. Pour chaque permutation, on permute les signatures bit à bit, on procède à un tri lexicographique de tous les éléments et on garde les $B$ plus proches éléments de l'élément requête dont l'approximation du cosinus est inférieur à un certain seuil. Cela évite d'avoir à calculer la distance de Hamming avec les $n$ autre éléments (complexité en $O(n^2)$ lorsque l'espace a autant de traits que d'éléments). L'algorithme effectue autant de tris que de permutations et a donc une complexité en $O(p.n.log(n)) \approx O(n.log(n))$.

Cet algorithme fonctionne car une signature permutée est une représentation valide du vecteur d'origine. En triant ces représentations par leur ordre lexicographique, on rapproche les signatures ayant une partie de leur signature similaire et donc une probabilité importante d'avoir une distance de Hamming faible. Le fait de produire $p$ permutations aléatoires permet de trouver différentes signatures ayant cette forte probabilité. Le calcul de la distance de Hamming est alors fait sur ces signatures et le tri est effectué sur un nombre négligeable d'éléments ($p.B$).

#### 3.2.1.2 Notre proposition

Soit $w_0$ le terme dont on cherche les plus proches voisins. D'une part, si on effectue un XOR de tous les éléments de la base avec le vecteur $w_0$, celui-ci obtient une représentation de bits exclusivement nuls et devient donc le premier élément de la liste si elle était trié lexicographiquement. On peut alors se permettre de ne pas trier les éléments mais de n'extraire que les $k$ premiers et de ne les trier eux-mêmes que par la suite. Cette opération a une complexité linéaire.

Nous procédons alors aux $p$ permutations et répétons l'opération d'extraction des $k$ premiers éléments pour chacune d'entre elle. On a donc une complexité en $O(n + kp.log(kp)) \approx O(n)$ à la place de $O(n.log(n))$.

Notre méthode est donc plus efficace si on souhaite rechercher les plus proches voisins d'un élément source au coup par coup. En revanche, s'il s'agit de rechercher les plus proches voisins de tous les éléments de la base, on doit répéter l'opération $n$ fois, et on a donc une complexité en $O(n^2)$ qui est supérieure à celle de l'algorithme modifié par [Ravichandran *et al.* 2005].

D'autre part, nous remarquons que le tri lexicographique sur un vecteur de bits prend en compte en moyenne deux bits, ce qui nécessite d'avoir un très grand nombre $p$ de permutations pour obtenir une bonne approximation. En effet, la probabilité qu'on ne prenne en compte que le premier bit est de $1/2$ (on a effectivement une chance sur deux que le premier bit soit identique), la probabilité pour qu'on prenne en compte deux et seulement deux bits est de $1/4$, la probabilité pour qu'on prenne en compte k et seulement k bits est de $(1/2)^k$. Le nombre moyen de bits pris en compte est donc l'espérance de la variable aléatoire de loi de probabilité $P_X(i) = (1/2)^i$. Cette espérance est donnée par la série $\sum_{k=1}^{n}(p_i.x_i) = \sum_{k=1}^{n}(k/2^k)$ convergeant vers 2 (voir encadré 3.9 pour la démonstration).

Soit $\sum_{k=0}^{\infty} x^k$ une série géométrique de raison x avec $|x| < 1$.
Par nature, elle converge vers $\frac{1}{1-x}$
On a donc :
$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$
En dérivant, on obtient :

$$\begin{aligned} \frac{1}{(1-x)^2} &= \sum_{k=1}^{\infty} kx^{k-1} \\ &= \frac{1}{x}\sum_{k=1}^{\infty} kx^k \\ \iff \sum_{k=1}^{\infty} kx^k &= \frac{x}{(1-x)^2} \end{aligned}$$

Si $x = \frac{1}{2}$ alors on obtient :

$$\boxed{\sum_{k=0}^{\infty} \frac{k}{2^k} = 2}$$

qed.

FIGURE 3.9 – Convergence de la série $\sum_{k=1}^{n} (k/2^k)$

C'est pourquoi, pour les $p$ permutations pour lesquelles on va extraire les $B$ plus proches voisins, on ne procède plus à une permutation bit à bit mais à une permutation de sous-parties de signature. Nous qualifions d'unitaires ces sous-parties. Pour chacune des signatures, le tri se fait sur le nombre de bits de chaque sous-partie unitaire. Pour exemple, si nous prenons des sous-parties unitaires de huit bits, on effectue une permutation sur ces sous-parties, puis un tri lexicographique sur des valeurs variant de 0 à 8. Le tri sur ces sous-parties unitaires conserve la propriété de rapprocher des vecteurs pour lesquelles la distance de Hamming est faible. En effet après le XOR, une sous-partie unitaire dont le compte de bits est égal à 0 correspond à une sous-partie identique à celle de la signature source.

Cela permettra de prendre en compte plus de bits lors du tri (la probabilité de prendre en compte les 8 bits de la 1ère sous-partie est de 1/2, les 16 bits de la première et deuxième sous partie 1/4, etc...) et donc de diminuer le nombre de permutations nécessaires à la précision de l'algorithme.

#### 3.2.1.3 Implémentation parallèle

L'implémentation parallèle que nous avons développée consiste à répartir des sous-parties de signatures (multiple des sous-parties unitaires que l'on vient de décrire) sur chaque processeur différent.

En effet, nous souhaitons que notre système intègre à la fois l'algorithme de recherche rapide et la procédure de hachage permettant d'introduire de nouveaux éléments dans l'espace. Pour cette procédure, la répartition optimale consiste à répartir les clés de hachage sur les différents processeurs pour que chaque processeur puisse calculer parallèlement une partie du hachage.

Avec une telle répartition, il est naturel de considérer que chaque processeur est responsable d'une partie des vecteurs hachés (plusieurs sous-parties unitaires). Chaque processeur s'occupe alors, pour chaque permutation, de l'extraction des $k$ meilleurs éléments qu'il envoie ensuite au processeur maître

pour être triés. Une répartition des permutations sur les processeurs n'aurait pas été plus efficace.

### 3.2.2 Différents pouvoirs de discrimination en fonction des espaces et des mots à discriminer

Grâce à l'algorithme décrit ci-dessus, nous avons pu calculer la liste des $k$ plus proches voisins de mots polysémiques pour chacun des espaces sémantiques (décrits en 3.1.1.2) et nous voyons par exemple dans le tableau 3.13 les résultats à k=10 pour les mots *barrage* et *vol* dans différents espaces. [13]

| | | |
|---|---|---|
| barrage | COMPDUNOM[1] | barrage, infrastructure, aménagement, amont, bâtiment, station, canal, installation, réacteur, parcours |
| | COD_V[2] | barrage, barrière, pont, bâtiment, chantier, centrale, usine, station, installation, banc |
| | APPOS[3] | barrage, pont, canal, empierrement, déchetterie, digue, viaduc, mousqueterie, raz, écluse |
| | APPOS.reverse[4] | barrage, digue, déversoir, pont, électrolyseur, redresseur, lanier, route, lac, autoroute |
| vol | COMPDUNOM | vol, avion, voyage, retour, course, achat, opération, première, journée, premier |
| | COD_V | vol, voyage, retour, création, attaque, changement, mise, mariage, acte, fin |
| | APPOS | vol, meurtre, racket, fraude, chantage, assassinat, homicide, rapine, violence, crime |
| | APPOS.reverse | vol, meurtre, viol, prostitution, rapine, adultère, proxénétisme, idolâtrie, agression, luxure |

[1] *complément du nom*

[2] *complément d'objet direct*

[3] *apposition*

[3] relation inverse de l'*apposition*

TABLE 3.13 – 10 plus proches voisins des mots *barrage* et *vol*

Pour un mot comme *barrage* dont les usages peuvent se décliner en : *construction sur un cours d'eau*, *infrastructure industrielle* et *obstacle*, les différents espaces retournent des listes de plus proches voisins où ces usages sont non différenciés. En revanche, on remarque que pour un mot comme *vol* les plus proches voisins obtenus sont assez différents selon les espaces utilisés. Par exemple les espaces *complément du nom* et *complément d'objet* mettent en valeur la proximité sémantique du *vol aérien* tandis que les espaces*apposition* et son espace inverse font référence au *délit de vol*. Ces différences s'expliquent par le fait que certains sens apparaissent beaucoup plus fréquemment que

13. Ces deux mots font partie de la liste des noms polysémiques étudiés dans la campagne d'évaluation Romanseval sur laquelle nous reviendrons lors de l'évaluation de notre désambiguïsateur.

d'autres dans une position syntaxique donnée. Ici, le *vol* en tant que délit apparemment beaucoup plus souvent de façon énumérative que le *vol* aérien, et l'inverse est également vrai pour les relations de *complément du nom* et *complément d'objet*. Les espaces contiennent donc des informations différentes que nous supposons utiles de garder distinctes.

### 3.2.3 Clustering de mots multi-représentés

Dans notre approche, nous cherchons à regrouper les plus proches voisins d'un mot source dans des ensembles représentant chacun un usage ou un sens de ce mot. On souhaite parvenir à distinguer différents clusters comme dans l'exemple manuel de la figure 3.10. Cette figure montre une projection à trois dimensions où les contextes devraient parvenir à discriminer trois significations du mot *barrage*.

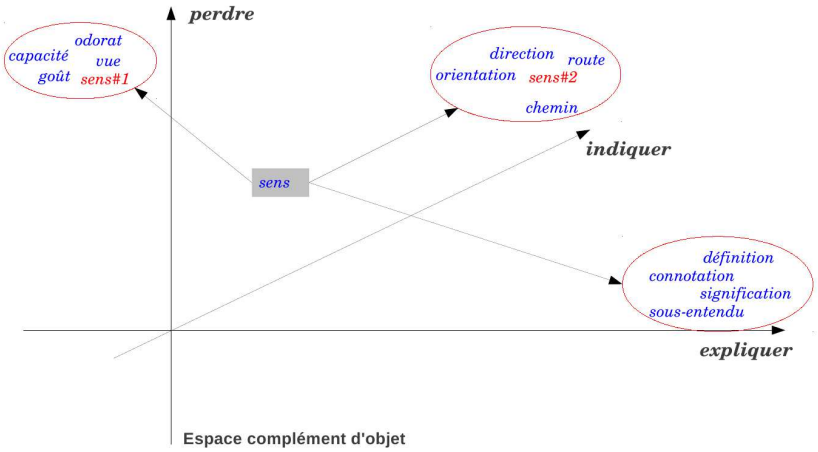



FIGURE 3.10 – Discrimination de sens pour le mot *sens* dans l'espace *complément_du_nom*

Dans cet exemple idéal, l'utilisation d'un unique espace vectoriel suffit. Mais on a vu plus haut que les différents espaces mettaient en relief différentes proximités. On voit par exemple dans le tableau 3.13, que l'espace COD_V met en relief la proximité sémantique sur le sens *déplacement* du mot *vol* alors que l'espace APPOS souligne la proximité avec le sens *délit*. Cette distinction n'étant pas systématique selon les espaces et les mots traités, nous souhaitons prendre en compte la spécificité de chacun des espaces tout en permettant le regroupement inter-espaces.

Pour cela, nous nous inspirons des algorithmes de clustering Shared Nearest Neighbours adaptés et utilisés par [Ertöz *et al.* 2001] et [Ferret 2004] ainsi que de l'algorithme HyperLex développé par [Véronis 2003]. Ces deux méthodes présentent l'avantage de ne pas avoir à fixer à l'avance le

nombre de clusters souhaité. Nous proposons deux adaptations de ces algorithmes pour obtenir deux méthodes de clustering multi-représenté par vote. On rappelle qu'un mot multi-représenté est un mot qui possède différentes signatures dans différents espaces, suivant ainsi la même définition que les objets multi-représentés de [Achtert *et al.* 2006]. Le clustering multi-représenté permet donc de tenir compte de ces différents espaces.

Pour chacune des parties du discours pour lesquelles on souhaite induire des sens (*substantif, verbe, adjectif*), on conserve les espaces

- pertinents à ces parties (par exemple lorsqu'on traite un verbe, on délaisse l'espace COMPDUNOM qui ne concerne que les substantifs et on utilise l'espace COD_V inverse et non l'espace COD_V qui est entièrement creux pour les verbes (cf. tableau 3.2))
- et correspondant à des relations entre mots pleins (substantif, verbe, adverbe, adjectif) (on délaisse par exemple la relation reliant un déterminant à son substantif).

De plus pour ne conserver que les espaces significatifs, on filtre ces espaces pertinents manuellement. Ce filtrage est nécessaire pour se défaire des espaces dont le bruit est trop grand dû à la rareté des phénomènes (c'est le cas notamment des relations rares du type attribut de l'objet, substantif juxtaposé...) ou aux erreurs de l'analyseur. Après visualisation, un espace est dit significatif lorsque les premiers plus proches voisins d'un ensemble de mots tests nous ont semblé être sémantiquement liés (*e.g. bateau : navire, véhicule, avion, bâtiment, train, camion, voiture, appareil*). Les espaces de fenêtre ont été exclus de cette liste, ceux-ci apportant trop de mots associés dont la sémantique est liée mais non similaire (*e.g. bateau : navire, mer, port, bord, endroit, voiture, loin, route, vent*). Ces mots ne sont pas mauvais en soi pour définir un cluster de sens mais ils risqueraient de provoquer du bruit lors de notre apprentissage. En effet, les éléments appartenant aux clusters étant considérés comme des données d'apprentissage, il est impératif que ceux-ci soit sémantiquement proche du mot ambigu à classifier. Les espaces significatifs conservés sont finalement l'ensemble suivant : { complément d'objet direct, complément du nom, sujet du verbe, apposition, adjectif épithète post-nominal inverse, adjectif épithète pré-nominal inverse}.

#### 3.2.3.1 Méthode inspirée de l'algorithme Shared Nearest Neighbours

**Méthode originale** L'algorithme original de [Ertöz *et al.* 2001] a déjà été employée par [Ferret 2004] pour l'induction de sens de mots. La méthode consiste tout d'abord à construire le graphe des plus proches voisins du mot source, où tous les mots sont représentés par des noeuds et les arcs prennent pour valeur la similarité cosinus entre ses deux mots extrémités dans l'espace sémantique considéré.

Un second graphe est ensuite construit, il est appelé graphe des plus proches voisins partagés. Dans un tel graphe, les arcs relient deux noeuds qui partagent au moins un voisin dans le premier

graphe. Ils prennent pour valeur la somme des voisins qu'ils partageaient dans le premier graphe.

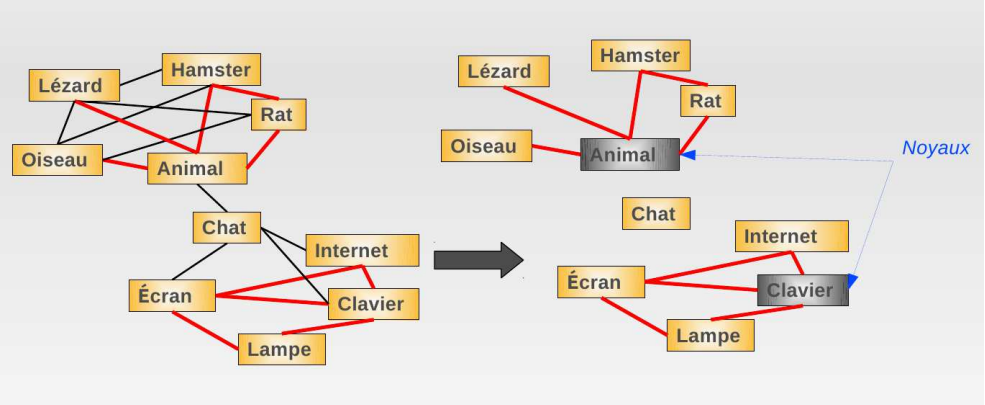



FIGURE 3.11 – Algorithme Shared Nearest Neighbours appliqué au clustering des coocurrents du mot *chat*

La suite de l'algorithme est illustré par la figure 3.11 représentant un exemple de discrimination de sens pour le mot *chat*. Un seuil dit de *lien fort* est fixé pour éliminer les arcs trop faibles auxquels on ne peut pas accorder une confiance suffisamment grande. Les arcs noirs sont donc éliminés du graphe. On calcule ensuite le degré des noeuds restants (le nombre d'arcs qu'ils possèdent). Les noeuds dont le degré dépasse un autre seuil sont alors déclarés noyaux des clusters à générer. C'est ici le cas de *animal* et *Internet*. Tous les autres éléments sont ensuite attribués au cluster avec lequel ils partagent le lien le plus fort et constituent ainsi des clusters de sens. Nous trouvons dans cet exemple le sens de *chat* en tant qu'animal ainsi que de *chat* pour l'emploi du mot anglais correspondant à une salle de discussion virtuelle. Ici le mot *chat* lui-même reste ambigu et n'appartient à aucun des clusters.

Enfin, les clusters dont les noyaux sont trop proches sont finalement joints et les éléments appartenant à des clusters jugés trop petits sont réassignés.

Ainsi, cette méthode permet de générer un certain nombre de clusters que l'on n'a pas besoin de fixer à l'avance, ce qui est indispensable dans le cadre de l'induction de sens, puisqu'on ne sait pas à l'avance de combien de sens un mot dispose.

**Notre proposition** La modification de l'algorithme que nous proposons tient compte du fait que les mots sont présents dans plusieurs représentations $R_i$ différentes, correspondant chacune à un des espaces sémantiques issu d'une relation syntaxique donnée.

Notre méthode reprend la méthode de choix de noyaux de l'algorithme SNN de [Ertöz *et al.* 2001] pour lequel on n'a pas à choisir le nombre de clusters *a priori*. Ainsi on constitue un graphe dans chacune des représentations distinctement.

L'expérimentation nous donnant de meilleurs résultats pour un graphe de voisins directs qu'avec le graphe des Plus Proches Voisins Partagés, nous garderons le premier. Une explication possible

concernant ce résultat surprenant comparé aux résultats exposés dans [Ertöz *et al.* 2001] est le fait que nous n'utilisons qu'une petite partie des éléments de l'espace d'origine, on a donc trop peu de données pour exploiter ce second degré d'information. Ayant ainsi perdu la notion de *lien fort* définie comme reliant deux nœuds partageant un nombre de voisins supérieur à un certain seuil, nous modifions notre critère de sélection de noyaux en conséquence. Le nom Shared Nearest Neighbours n'ayant plus de raison d'être, nous nous référerons désormais à cet algorithme sous le nom de MultiNN.

Nous disposons donc d'un graphe de plus proches voisins pour chaque représentation $R_i$. La moyenne des distances des $n$ plus proches voisins étant très différente d'une représentation à une autre, nous rendons auto-adaptatifs tous les seuils utilisés dans l'algorithme de la façon suivante.

Soit l'ensemble $E$ des éléments $e_m$ à clusteriser. $x_i(e)$ la variable à seuiller dans $R_i$. On fixe un seul paramètre $param\_quota$ commun à tous les espaces. On définit alors $seuil_{x_i}$ de la façon suivante :

$$seuil_{x_i} = \min_{e \in E} x_i(e) + param\_quota.(\max_{e \in E} x_i(e) - \min_{e \in E} x_i(e))$$

Ceci est illustré par la figure 3.12.

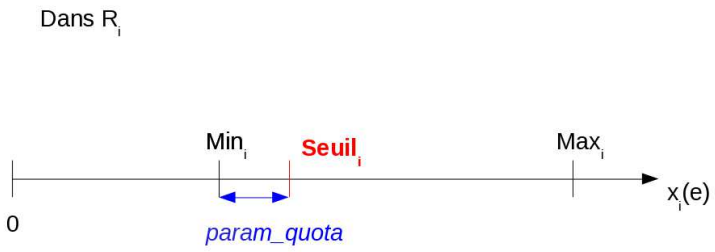



FIGURE 3.12 – Seuil auto-adaptatif dans $R_i$

L'algorithme modifié se déroule de façon assez similaire à l'algorithme d'origine, tout en prenant en compte chacun des espaces pour prendre les différentes décisions par vote majoritaire. Ainsi, dans chacun des espaces significatifs (COD_V, SUJ_V...) :

1. on établit le graphe des plus proches voisins dans lequel les noeuds sont les éléments de E et les arcs relient tous les noeuds entre eux et ont pour valeur la similarité cosinus approchée (LSH) que l'on calcule entre les deux noeuds extrémités ;
2. on définit un seuil de lien fort et on élimine les arcs dont les valeurs sont inférieures à ce seuil (paramètre noté $seuilFaiblesse$) ;
3. pour chaque noeud du graphe on calcule la somme des valeurs de ses liens :
4. si cette somme est plus grande qu'un certain seuil (paramètre noté $seuilNoyau$), on dit que ce point est un *noyau local* pour cet espace.

5. si cette somme est plus petite qu'un certain seuil (paramètre *seuilBruit*), on dit que ce point est un *bruit local* pour cet espace.

Pour tout élément de E, on sait le nombre d'espaces dans lesquels il est *noyau local*. Si ce nombre est supérieur à certain pourcentage des espaces utilisés (passé en paramètre *ratio*), cet élément est dit *noyau global*. On détermine de la même façon les éléments qui s'avèrent être *bruit global*, respectivement à la notion de *bruit local*. Ces éléments sont retirés de $E$.

Nous constituons ensuite des clusters autour de chaque noyau global de la façon suivante. Dans chaque espace, on retire de E les éléments qui ont été désignés *noyaux globaux*. Puis, dans chacun des espaces et pour chaque élément de E, on enregistre un vote pour le noyau global le plus proche (cosinus approché le plus élevé). On somme les votes sur tous les espaces et on assigne chaque élément à son noyau le plus populaire (éventuellement à plusieurs en cas d'égalité).

Dans chaque espace, si la distance entre deux noyaux globaux a une valeur inférieure à un seuil donné (noté *seuilJointure*, l'espace vote pour la jonction des deux clusters associés. Si un pourcentage suffisant d'espaces votent pour cette jonction, les deux clusters sont regroupés (le même paramètre *ratio* est utilisé).

Les éléments des clusters considérés trop petits par rapport au nombre d'éléments à regrouper sont réassignés un à un aux gros clusters.

#### 3.2.3.2 Méthode fondée sur l'algorithme HyperLex

Nous adaptons une seconde méthode à notre multi-représentation des mots. Il s'agit de l'algorithme HyperLex présenté par [Véronis 2003]. Celui-ci est à l'origine appliqué à une liste de cooccurrents mais dans notre cas nous l'appliquons à une liste de plus proches voisins.

Nous avons remarqué par l'analyse manuelle de ces listes que certains sens ne voient pas (ou très peu) leur vocabulaire apparaître dans les plus proches voisins du mot source. Cette remarque est valable même lorsqu'il s'agit d'un sens principal et non d'un usage, notamment quand un sens et beaucoup plus fréquent qu'un autre dans le corpus. C'est par exemple le cas pour le mot *avocat*. Il faut sinon remonter trop loin dans les plus proches voisins et induire alors beaucoup trop de mots bruit. Nous avons également essayé d'appliquer un filtrage consistant à ne garder un mot que s'il était présent dans les plus proches voisins d'au moins deux des espaces, mais ce filtrage a tendance à exclure les mots qui nous intéressent et garder un certain nombre de mots bruités.

Pour pallier ce problème, nous avons enrichi nos listes de mots à clusteriser de la façon suivante. Nous avons produit pour chaque mot source une autre liste issue des meilleurs cooccurrents du second ordre. Pour chaque mot source dans chaque espace significatif, nous avons extrait les 50 cooccurrents dont l'information mutuelle avec le mot source était la plus élevée. Nous avons alors calculé dans les espaces syntaxiques inverses les 10 meilleurs cooccurrents des 50 cooccurrents de premier ordre. Les cooccurrents de second ordre apparaissant au moins trois fois dans la liste finale ont été conservés. Nous avons ensuite fait l'union de cette liste filtrée avec les 10 plus proches voisins

du mots source apparaissant dans chacun des espaces significatifs. Ces listes montrent une variété plus importante de lexique lié aux mots sources.

**Méthode originale** L'algorithme original est le suivant. Soit E l'ensemble des mots à regrouper. Pour chacun d'entre eux, on calcule la distance au mot source. Le mot le plus proche devient noyau d'un cluster. On attribue à ce cluster tous les éléments appartenant à la sphère de rayon $\phi$. Si leur nombre est supérieur à un paramètre $m$, le cluster est valide et on retire ces éléments de E. Sinon on remet le noyau dans l'ensemble E et on continue. Le mot suivant de E le plus proche du mot source devient noyau d'un autre cluster. On réitère jusqu'à ce que E soit vide.

**Nos propositions** Avec nos multiples représentations, nous utilisons trois modes différents pour le choix successif des noyaux. Le premier consiste à considérer les distances des éléments de l'ensemble E par rapport au mot source comme la moyenne des distances dans tous les espaces. Le deuxième consiste à prendre le minimum des distances et le troisième à prendre le maximum.

Nous proposons en outre une deuxième approche distincte de ce qui est proposé dans l'algorithme original. Supposons que si, pour un élément donné, dans un espace donné, la variance de sa distance aux autres éléments est élevée, cela signifie que dans cet espace il est un bon discriminateur de sens, certains mots étant très proches de lui et d'autres très lointains. Il serait alors un bon élément pour être noyau d'un cluster. À partir de cette hypothèse, on peut alors choisir le noyau suivant comme étant l'élément ayant la moyenne de ces variances sur les différentes espaces la plus élevée, ou encore le minimum ou le maximum de ces variances.

Ces six heuristiques seront testées lors de la phase d'évaluation au chapitre suivant.

Pour cet algorithme également et pour les raisons invoquées précédemment (les distances d'un espace à un autre ne sont pas comparables), nous avons rendu auto-adaptatif le seuil de validation d'attribution d'un élément à un cluster. Le seuil $\phi$ est également dépendant du mot noyau auquel vont se rattacher les différents éléments et est donné par la formule suivante. Soit $\mathcal{R}$ l'ensemble des représentations.

$$\forall R_i \in \mathcal{R}, \phi_i(noyau) = \frac{1}{|E|}.\sum_{e \in E} distance(noyau, e) - \alpha.\sqrt{\frac{1}{|E|}.\sum \sigma_i^2(noyau)}$$

avec $\alpha$ un paramètre donné en entrée et $\sigma_i^2$ la variance des distances du *noyau* à tous les éléments de $E$ dans l'espace $R_i$.

Un espace $R_i$ vote alors pour une attribution de $e$ au cluster $c$ si la distance séparant $e$ du noyau de $c$ est inférieure à $\phi_i(noyau)$. Au final, les éléments sont attribués au cluster uniquement si un certain pourcentage d'espaces vote pour cette attribution (*ratio* passé en paramètre).

Enfin, les gros clusters comportant presque systématiquement des sens mélangés, nous proposons d'utiliser un seuil maximal (noté $M$)pour la taille des clusters, au delà de laquelle une ré-attribution de chacun des éléments des clusters incriminés est effectuée en fin de traitement. Chaque espace vote pour le cluster dont la moyenne des distances entre le mot et les différents éléments du cluster est la plus faible. Les attributions supportées par un nombre de votes supérieur au ratio sont validées. On réitère cette phase en augmentant le seuil maximal à chaque itération. Ce pas d'augmentation est aussi passé en paramètre (noté $p$). Ces itérations s'arrêtent quand le nombre total de clusters est égal à 3.

### 3.2.4 Résultats

Les résultats présentés ici ont été obtenus empiriquement sans optimisation automatique des paramètres dans cette première étape. Pour l'exemple du mot *barrage*, nous disposons des résultats fournis par les méthodes dont nous nous sommes inspirées.

Nous présentons ainsi nos résultats pour le mot *barrage* dans le tableau 3.14. Nous avons privilégié un grand nombre d'éléments à regrouper afin de rassembler le maximum de sens possibles. Nous avons aussi opté pour un résultat contenant un grand nombre de clusters afin d'en obtenir le maximum possible de précis, quitte à en obtenir certains qui pourraient encore être regroupés. Un classifieur destiné à faire de la désambiguïsation et apprenant sur ces données saura ne pas classifier de mot dans l'une de ces classes si les éléments de celle-ci sont trop clairsemés et ont des équivalents dans d'autres clusters.

La comparaison de nos clusters avec ceux des algorithmes d'origine reste difficile puisque nous ne cherchons pas à regrouper le même type de terme (cooccurrents vs. plus proches voisins syntaxiques). Cependant nous pouvons voir que les sens distingués ne sont pas toujours les mêmes. Nous retrouvons d'une part en 3.1, 3.4 et d'autre part en 4.3 les sens de *barrage* correspondant au *barrage frontalier, policier ou routier*, que l'on ne distingue pas bien entre eux. Le Petit Larousse utilisé dans la campagne ROMANSEVAL [Segond 2000] ne fait lui-même pas la distinction et considère simplement ceux-ci comme *obstacle*. En 3.5, 3.6, et 4.5, l'usage du *barrage hydraulique* est distingué en tant qu'*infrastructure industrielle*, en 3.3 et 4.6 et 4.10 en tant que *construction sur un cours d'eau*. On trouve également en 4.2 et 4.4 sa présence en tant que *générateur d'électricité* ou en 4.6 en tant qu'*élément du paysage*. Il s'agit du même objet physique mais l'usage est différent. En revanche on ne retrouve pas le sens du *match de barrage* pour lequel il existe très peu de mots partageant les mêmes contextes syntaxiques. Pour le mot *vol*, notre algorithme extrait correctement les sens *délit de vol* et *vol aérien*, ce qui n'était pas le cas pour l'algorithme HyperLex.

Globalement, l'algorithme MultiHyperLex semble donner des meilleurs résultats que l'algorithme MultiNN. Cela reste à confirmer par une évaluation automatique à laquelle nous procéderons lors de la phase de désambiguïsation. Celle-ci permettra d'une part d'optimiser les différents paramètres et d'autre part d'évaluer la qualité de discrimination des clusters obtenus.

| Mot-source | barrage |
|---|---|
| HyperLex [Véronis 2003] | 1.1 : eau, construction, ouvrage, rivière, projet, retenue, crue<br>1.2 : routier, véhicule, camion, membre, conducteur, policier, groupement<br>1.3 : frontière, Algérie, militaire, efficacité, armée, Suisse, poste<br>1.4 : match, vainqueur, victoire, rencontre, qualification, tir, football |
| SNN [Ferret 2004] | 2.1 : manifestant, forces_de_l'ordre, préfecture, agriculteur, protester, incendier, calme, pierre<br>2.2 : conducteur, routier, véhicule, poids_lourd, camion, permis, trafic, bloquer, voiture, autoroute<br>2.3 : fleuve, lac, rivière, bassin, mètre_cube, crue, amont, pollution, affluent, saumon, poisson<br>2.4 : blessé, casque_bleu, soldat, milicien, tir, milice, convoi, évacuer, croate, milicien, combattant |
| MultiNN | 3.1 : **accès, entrée, obstacle, parcours**<br>3.2 : aménagement, commune, emplacement, infrastructure<br>3.3 : **bâtiment, bassin, chantier, emplacement, parc, pont, tunnel**<br>3.4 : **dispositif, commune, résistance**<br>3.5 : **installation, barrière, centrale, chaîne, commune, dépôt, emplacement, fondation, infrastructure, réserve, usine**<br>3.6 : **station, canal, centrale, chantier, commune, emplacement**<br>3.7 : **transport, commune, dépôt, distribution** |
| MultiHyperLex | 4.1 : armoire, pare-brise, fourche<br>4.2 : **aval, modulation, amont**<br>4.3 : **colonnade, parpaing, télescope, feu, fond, parcours, vallon, butte, mausolée**<br>4.4 : **électrification, étiage, cogénération**<br>4.5 : **oléoduc, turbine, captage**<br>4.6 : **oued, djebel, rizière**<br>4.7 : ponceau, talus, déversoir<br>4.8 : poutrelle, ancre, gravier<br>4.9 : raz, bulldozer, orage<br>4.10 : **terril, sédiment, estuaire** |

TABLE 3.14 – Comparatif des clusters construits à partir du mot *barrage*

Une évaluation plus pertinente de ces différents types de clusters serait de les utiliser dans une tâche applicative (désambiguïsation de sens ou recherche d'information) et d'évaluer les résultats de celle-ci. Dans le cadre de l'étude du désambiguïsateur que nous présenterons dans le chapitre suivant, nous emploierons notamment la V-mesure définie par [Rosenberg & Hirschberg 2007] afin d'évaluer la qualité de discrimination des clusters produits.

### 3.2.5 Premières conclusions, discussions et perspectives

Un avantage des méthodes présentées peut être mis en avant : nous supposons que le fait d'utiliser les plus proches voisins comme ensemble d'éléments à regrouper (et non les cooccurrents), permet d'utiliser ces voisins comme données d'apprentissage d'un classifieur de désambiguïsation lexicale. La suite de ce manuscrit montre les résultats d'une telle expérimentation.

Il serait intéressant de comparer les clusters que nous avons obtenu par combinaison des espaces avec des clusters obtenus à partir des mêmes ensembles de plus proches voisins que dans cette étude mais en utilisant d'une part l'espace de cooccurrences de fenêtres seul ainsi que chacun des espaces syntaxiques seul, et en utilisant d'autre part les matrices concaténées.

Nous aurions également pu faire état d'une comparaison des ressources obtenues par l'induction avec des ressources manuelles existantes en utilisant par exemple des mesures de précision et de rappel de recouvrement de synsets entre nos clusters et les synsets d'un WordNet français de référence. Cependant cela nous pose deux problèmes essentiels. D'une part les ressources manuelles ne détiennent pas nécessairement une vérité absolue en terme de distinctions de sens. En effet, les besoins en terme de granularité de distinction de sens, ainsi que de couverture des différents sens et expressions figées (phrasèmes), varient très fortement d'une application à une autre ([Véronis 2004]). D'autre part, comme nous l'avons montré dans la première partie de ce chapitre, on ne peut pas vraiment dire qu'il existe de ressource de référence WordNet en français.

Enfin, les algorithmes de désambiguïsation de sens sont rarement indépendants de la ressource qu'ils exploitent pour procéder à la désambiguïsation. Ainsi on peut considérer que la qualité d'un système de désambiguïsation dépend à la fois des ressources qu'il exploite et de l'algorithme qui le fait fonctionner. L'évaluation de la qualité de nos sens passera donc par la tâche de désambiguïsation elle-même et sera ainsi évaluée conjointement.

## 3.3 Bilan

La ressource WordNet française que nous avons obtenue par combinaison de ressources manuelles et automatiques est d'une couverture bien supérieure à ce que l'on pouvait trouver jusqu'à présent parmi les ressources de langue française. La perte en précision n'est pas inexistante mais elle reste raisonnable et nous espérons que cette ressource pourra servir à la communauté en l'état.

Par ailleurs, nous disposons désormais d'un répertoire de sens induits de façon complètement automatique. Le chapitre suivant nous permettra conjointement de mettre en œuvre un désambiguïsateur lexical utilisant ce répertoire, d'optimiser les paramètres de l'induction et d'évaluer la qualité des clusters optimisés.

# Chapitre 4

# Les espaces distributionnels au cœur de la désambiguïsation lexicale

Comme nous l'avons vu dans notre état de l'art (Section 2.4.1), le problème de la désambiguïsation lexicale est loin d'être nouveau et de multiples solutions supervisées sont déjà effectives. Nous nous heurtons néanmoins à un problème assez récurrent, il n'existe que très peu de données pour le français.

Nous nous intéressons ainsi à la possibilité de mise en œuvre d'un système non-supervisé, au sens où il ne nécessite pas de corpus d'apprentissage sauf pour fixer quelques paramètres. Nous décrivons dans un premier temps ce système exploitant nos sens de mots induits. Nous étudions ensuite la qualité de désambiguïsation de notre algorithme par une évaluation automatique à l'aide des données de la campagne Romanseval ([Segond 2000]). Enfin, nous terminons ce chapitre par des recommandations pour l'intégration d'un tel module dans un système de recherche de documents.

## 4.1 Description du système proposé

Nous rappelons les objectifs de la désambiguïsation lexicale tels que nous les avons définis en introduction de ce manuscrit. Le système dispose d'une ressource lexicale de référence dans laquelle un vocabulaire est répertorié. Pour chaque mot de ce vocabulaire, la ressource distingue ou non plusieurs sens possibles. [1]

Notre système d'analyse sémantique prend en entrée le résultat d'une analyse morphosyntaxique de texte. Chaque lemme de ce texte possédant plusieurs sens dans la ressource lexicale est un lemme ambigu. Le cas échéant, le système doit alors déterminer auquel de ces sens le lemme correspond.

1. Nous ne reviendrons pas ici sur la granularité de cette distinction ni sur son aspect discret ou continu (pour plus de détails sur le sujet cf. 2.2.2).

### 4.1.1 Description fonctionnelle

L'inventaire de sens utilisé par notre système est un ensemble de clusters de mots construit par les méthodes décrites à la section 3.2. Pour chaque mot $m$ du vocabulaire, on soumet au *clustering* un ensemble de mots sélectionnés parmi les plus proches voisins et les cooccurrents du second ordre de $m$. Chacun des clusters ainsi constitué représente un sens.

Pour chaque nouvelle occurrence de $m$ dans un texte, le système utilise l'analyse morphosyntaxique pour déterminer un vecteur caractéristique dans chacun des espaces sémantiques. Les mots appartenant aux clusters possèdent également un vecteur caractéristique dans chacun des espaces sémantiques. Bien que ces vecteurs caractéristiques ne soient pas calculés sur une phrase précise mais sur un corpus entier, on peut les considérer comme des données d'apprentissage du classifieur. En effet, de par le choix des mots à regrouper en clusters, leurs vecteurs caractéristiques correspondent aux modèles de distribution des mots sémantiquement les plus proches du mot $m$ et dont le sens est désambiguïsé par le clustering.

Dès lors, on pourra considérer que la tâche de désambiguïsation est en réalité une tâche de classification pour laquelle chaque occurrence du mot $m$ dans une phrase est l'élément à classifier et les différents clusters sont les différentes classes auxquelles le mot $m$ peut être attribué.

Nous avons montré à la section 3.2.2 que les espaces sémantiques utilisés possèdent un pouvoir de discrimination différent en fonction des espaces et des mots à discriminer, nous avons donc choisi de tenir compte de ces différences en utilisant un classifieur multi-représenté comme celui de [Kriegel *et al.* 2005].

### 4.1.2 Implémentation

Nous décrivons ci-après la méthode que nous proposons. Nous prendrons l'extrait de texte suivant pour illustrer notre propos :
*Ces génies musicaux utilisent aussi leurs propres* ***organes*** *vocaux. Souvent, ils confient leurs pensées en paroles plutôt qu'en notes.*

Nous nous concentrons sur le mot *organe* qui fait partie de la liste des mots polysémiques étudiés dans Romanseval.

#### 4.1.2.1 Inventaire de sens

L'inventaire de sens utilisé par notre système de désambiguïsation est constitué d'après la description faite à la section 3.2. On a donc pour chaque mot source du vocabulaire plusieurs clusters de mots sémantiquement similaires correspondant à différents usages du mot source. Par exemple, pour le mot *organe* on dispose des deux clusters de mots figurant dans le tableau 4.1. Le premier

cluster fait référence à *un organe administratif, judiciaire ou institutionnel* tandis que le deuxième cluster fait référence à *l'organe appartenant à un corps vivant.*

| **organe** | |
|---|---|
| autorité | poumon |
| gestion | peau |
| personnel | cellule |
| tribunal | organisme |
| droit | foie |
| comité | rein |
| membre | nerf |
| décision | cerveau |
| ministère | muscle |
| communauté | |
| institution | |
| organisation | |
| commission | |

TABLE 4.1 – Deux clusters de sens pour le mot *organe*

#### 4.1.2.2 Classification

Dans une phrase donnée $\phi$, tout nouveau terme $t$ pour lequel on possède plusieurs clusters dans notre inventaire de sens est considéré comme ambigu et est à classifier. Nous définissons ainsi $TC(\phi)$ l'ensemble des éléments de la phrase $\phi$ que l'on doit classifier. On a donc :

$$TC(\phi) = \{t \in termes(\phi), |clusters(lemme(t))| > 1\}$$

où $clusters(l_x)$ est l'ensemble des clusters induits pour le lemme $l_x$.

$t$ peut alors être représenté dans les espaces sémantiques utilisés pour la constitution des clusters de sens. En effet, la représentation vectorielle de $t$ dans les espaces $R_i$ constitués à partir de fenêtre de taille fixe est directe. Les coordonnées du vecteur $\overrightarrow{t}_{R_i}$ sont incrémentées si elles correspondent aux mots se trouvant dans le contexte fenêtre de $t$ dans la phrase $\phi$, et le vecteur est normalisé pour donner la représentation de $t$ dans $R_i$.

Reprenons notre exemple pour illustrer la constitution de nos vecteurs caractéristiques :
*Ces génies$_{-5}$ musicaux$_{-4}$ utilisent$_{-3}$ aussi$_{-2}$ leurs propres$_{-1}$* ***organes*** *vocaux$_1$. Souvent$_2$, ils confient$_3$ leurs pensées$_4$ en paroles$_5$ plutôt$_6$ qu'en notes$_7$.*

Dans la représentation issue des contextes fenêtres de taille [-5 ; +5 ], on aura le vecteur représentatif suivant :

| | génie | note | parole | pensée | (...) | musical | propre | vocal | (...) |
|---|---|---|---|---|---|---|---|---|---|
| $\overrightarrow{t}_{R_{window[-5;+5]}} = [$ | 0.1 | 0 | 0.1 | 0.1 | (0) | 0.1 | 0.1 | 0.1 | (0) |

| confier | utiliser | (...) | aussi | plutôt | souvent | (...) | |
|---|---|---|---|---|---|---|---|
| 0.1 | 0.1 | (0) | 0.1 | 0 | 0.1 | (0) | ] |

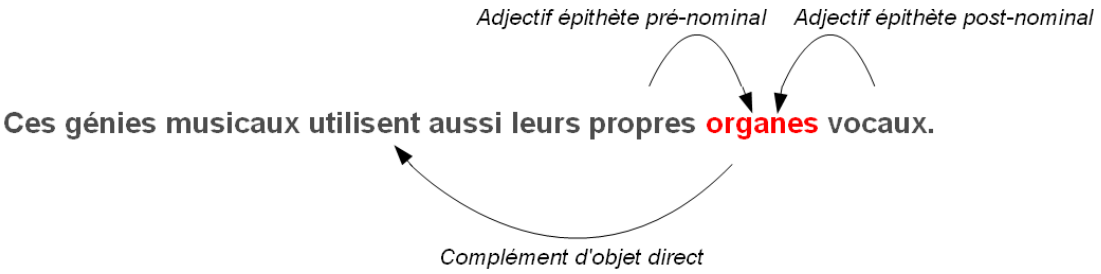



FIGURE 4.1 – Dépendances syntaxiques du terme *organe* dans la phrase exemple

Pour les représentations issues de l'analyse syntaxique, le procédé est assez semblable. L'exemple d'analyse syntaxique de la figure 4.1 nous donnera les vecteurs représentatifs correspondants : le mot *organe* apparaît une fois au contexte syntaxique de l'adjectif post-nominal *vocal*, la cellule correspondante du vecteur représentatif dans l'espace *adjectif post-nominal* est donc incrémentée et le procédé est répété pour les autres relations syntaxiques. On obtient donc les représentations suivantes du mot *organe* dans la phrase exemple :

| **Analyse syntaxique** | | **Vecteurs caractéristiques** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | propre | vocal | (...) | utiliser | (...) | |
| adj_postnominal(organe, vocal) | | $\overrightarrow{t}_{R_{adj\_postnominal}}$ = [ | 0 | 1 | (0) | 0 | (0) | ] |
| adj_prenominal(organe, propre) | => | $\overrightarrow{t}_{R_{adj\_prenominal}}$ = [ | 1 | 0 | (0) | 0 | (0) | ] |
| objetdirect(utiliser, organe) | | $\overrightarrow{t}_{R_{objetdirect.inverse}}$ = [ | 0 | 0 | (0) | 1 | (0) | ] |

Une fois ces vecteurs représentatifs calculés, il reste à classifier l'occurrence ambiguë parmi les différentes classes que sont les clusters de sens. Soit $\mathbf{C}=\{C_j\}$ l'ensemble de ces classes. Disposant de plusieurs représentations contenant des informations complémentaires, nous appliquons le classifieur k-NN adapté aux multiples représentations décrit par [Kriegel *et al.* 2005] afin d'exploiter l'information la plus fiable de chaque espace. Nous rapportons ici les détails de cette méthode.

Soient $o$ l'occurrence à classifier, et $\mathbf{R} = \{R_i\}$ l'ensemble des représentations dans lesquelles $o$ possède une caractérisation. $\forall i \in [|1, |R||]$, on note $o_i$ le vecteur caractéristique de $o$ dans $R_i$.

Soit $S_i(o,k)$ la sphère de centre $o_i$ et de rayon $d_i(o_i, nn_i(o,k))$ avec $nn_i(o,k)$ le $k$-ième plus proche voisin de $o_i$ dans $R_i$. Dans notre cas, on utilise : $d_i(o_i, nn_i(o,k)) = 1 - cos(o_i, nn_i(k))$.

Le classifieur permet de combiner les sphères de voisinages $S_i(o,k)$ de toutes les représentations tout en tenant compte de leur qualité. La notion de qualité est donnée par l'idée qu'une sphère de faible entropie, c'est-à-dire dont le voisinage contient des éléments appartenant à un petit nombre de classes différentes mais avec beaucoup d'éléments appartenant à chaque classe est plus fiable qu'une

sphère de forte entropie, c'est-à-dire une sphère contenant des éléments appartenant à un grand nombre de classes diverses avec chacune peu d'éléments dans $S_i(o,k)$.

Pour modéliser cette confiance, [Kriegel *et al.* 2005] définissent $cv(o)$ un vecteur de confiance (*confidence vector*, noté $cv$) associé à tout objet $o$. $cv_{i,j}(o)$ donne pour chaque représentation $R_i$ la confiance que l'on peut avoir dans la classification à la classe $C_j$. La définition de $cv(o)$ est donnée par la formule suivante :

$$\forall j, 1 \leq j \leq |C| : \quad (A) \quad cv_{i,j}(o) = \frac{\sum_{u \in S_i(o,k) \wedge c(u)=c_j} \frac{1}{d_{norm}(o,u)^2}}{\sum_{k=1}^{|c|} cv_{i,k}(o)} \tag{4.1}$$

$$\text{avec } d_{norm}(o,u) = \frac{d_i(o,u)}{max_{v \in S_i(o,k)} d_i(o,v)}.$$

Nous remarquons que le vecteur de confiance tel que défini ci-dessus dépend du nombre d'éléments dans la sphère de voisinage $S_i(o,k)$ appartenant à la classe $c_j$ donnée $(|u \in S_i(o,k)) \wedge c(u) = c_j|)$ mais qu'il ne tient pas compte du nombre d'éléments d'apprentissage de chacune de ces classes $c_j$. Or, d'après nous il est plus probable qu'un élément appartienne à une classe donnée si cette classe contient plus d'éléments que les autres. Ainsi nous proposons de tenir compte également du nombre d'éléments appartenant à chaque classe et testons également les trois formules suivantes :

$$(B) \quad cvB_{i,j}(o) = \frac{cv_{i,j}}{|c_j|}(o) \tag{4.2}$$

$$(C) \quad cvC_{i,j}(o) = \frac{cv_{i,j}}{log(1+|c_j|)}(o) \tag{4.3}$$

$$(D) \quad cvD_{i,j}(o) = \frac{cv_{i,j}}{log(10+|c_j|)}(o) \tag{4.4}$$

Nous utilisons alors le paramètre noté *modeCV* pour déterminer si la formule employée est $A, B, C$ ou $D$.

La désambiguïsation est ensuite effectuée en assignant à chaque mot ambigu la classe résultante de la règle de classification suivante :

$$Cl_{mr}(o) = argmax_{j=1,\ldots,|C|}(\sum_{i=1}^{|R|} w_i.cv_{i,j}(o)) \tag{4.5}$$

avec $w_i$ l'entropie de la représentation $R_i$ par rapport à toutes les classes possibles. Ce terme permet de pondérer la confiance accordée à chaque classe pour chaque représentation par la confiance générale accordée à une représentation. Il est défini par [Kriegel *et al.* 2005] de la façon suivante :

$$w_i = \begin{cases} 0 & si\ \phi_i(o) = \emptyset \\ \dfrac{1 + \sum_{j=1}^{|C|}(cv_{i,j}(o).log_{|C|}cv_{i,j}(o))}{\sum_{k=1}^{|R|}(1 + \sum_{j=1}^{|C|}(cv_{i,j}(o).log_{|C|}cv_{i,j}(o)))} & sinon \end{cases} \quad (4.6)$$

Enfin, pour la classification, tous les espaces concernant les noms sont exploitables si la relation est rencontrée dans le contexte du mot ambigu, excepté l'espace *Attribut de l'objet* trop bruité, soit l'ensemble : {*apposition, apposition inverse, complément d'objet, complément du nom, complément du nom inverse, adjectif épithète pré-nominal inverse, adjectif post-nominal inverse, sujet du verbe, modificateur du nom inverse, adverbe modifiant un nom inverse, complément de l'adjectif, attribut du sujet inverse, fenêtre de taille 5, fenêtre de taille 10*}.

## 4.2 Évaluation

Nous présentons ici l'évaluation de notre système de désambiguïsation des noms utilisant les clusters de mots induits.

Nous présentons dans un premier temps le protocole expérimental mis en place et analysons nos résultats par comparaison avec les données issues de la campagne Romanseval. Dans un deuxième temps, nous rapprochons nos résultats de la campagne d'évaluation SemEval sur la langue anglaise, celle-ci proposant une tâche spécifique à l'induction de sens de mots.

### 4.2.1 Campagne d'évaluation ROMANSEVAL

ROMANSEVAL est une campagne de désambiguïsation lexicale destinée au traitement des langues latines. La tâche dédiée à la langue française est décrite par [Segond 2000]. Pour le français, 60 mots ambigus (dont un tiers de noms, un tiers de verbes et un tiers d'adjectifs) sont annotés manuellement sur un corpus d'un million de mots pour un total de 3 724 entités annotées. Les annotations correspondent aux différents sens donnés par le dictionnaire *Petit Larousse*.

Dans cette sous-section, nous commençons par décrire les détails du protocole expérimental utilisé, puis nous décrivons notre algorithme de mapping. Nous procédons ensuite à l'optimisation des paramètres par la validation croisée de nos résultats. Enfin, nous présentons la performance globale de nos systèmes de désambiguïsation de noms et procédons à l'analyse de ces résultats.

#### 4.2.1.1 Protocole expérimental

Dans le protocole proposé lors de la campagne Romanseval, l'évaluation porte sur la qualité des annotations sous forme d'étiquettes de sens fournies par le référentiel de sens qu'est le dictionnaire

Petit Larousse.

Les sens produits par notre algorithme d'induction automatique doivent donc être mis en correspondance automatiquement avec les sens du dictionnaire. C'est ce que nous appelons la phase de *mapping*. Cette phase présente de grosses difficultés. En effet, les définitions fournies par ROMANSEVAL sont très hétérogènes : les définitions sont mélangées aux exemples sans distinction ce qui rend la tâche de mapping difficile.

Après désambiguïsation automatique du corpus ROMANSEVAL par notre système et notre inventaire de sens, chaque mot ambigu est annoté automatiquement avec son sens ROMANSEVAL par l'intermédiaire du mapping pré-calculé et les différentes mesures peuvent alors être effectuées.

Le protocole ainsi défini évalue à la fois la phase d'induction de sens préalable à la tâche de désambiguïsation, le mapping s'il y a lieu, ainsi que la désambiguïsation lexicale elle-même.

Parmi les différentes mesures proposées par [Segond 2000], nous utilisons la précision, le rappel et la F-mesure. Ces mesures sont calculées sur la tâche *fine-grained* où toutes les distinctions de sens les plus fines sont prises en compte. Les données de la campagne nous permettent d'une part d'étudier l'influence des différents paramètres, d'autre part de nous comparer à des systèmes français. La campagne d'évaluation n'est pas récente mais nous n'avons pas trouvé d'autres données françaises auxquelles nous comparer.

#### 4.2.1.2 Mapping

L'évaluation mise en place lors de la campagne ne distingue pas les systèmes utilisant un répertoire de sens externe aux données de la campagne et ne propose donc pas de données d'apprentissage pour apprendre la mise en correspondance des répertoires de sens (*mapping*) nécessaire à la résolution de la tâche.

Les participants n'ayant pas disposé de corpus d'apprentissage pour apprendre le mapping entre le référentiel de sens de leur système et les sens du Petit Larousse, nous ne procéderons pas non plus à un apprentissage supervisé de ces sens.

Commençons par observer le cas des trois définitions suivantes parmi celles fournies par Romanseval :

**constitution.n.I** action de constituer (quelque chose) ; ce qui en résulte, qui constitue (quelque chose) ; les éléments qui font partie d'un tout.
constitution d'un dossier, d'un gouvernement.

**constitution.n.III.1** droit
acte par lequel quelque chose est établi, constitué.
constitution d'une dot, d'une rente .

**constitution.n.IV** (avec une majuscule). ensemble des lois fondamentales qui établissent la forme d'un gouvernement, règlent les rapports entre gouvernants et gouvernés et déterminent l'organisation des pouvoirs publics.

Pour un cluster donné, l'idée générale du mapping consiste à attribuer un score de correspondance plus important aux définitions Romanseval dont les mots seraient souvent en présence des mots du cluster dans un texte général, externe au corpus donné. Nous faisons donc à nouveau intervenir nos matrices de calcul des espaces sémantiques, puisqu'elles contiennent entre autres les informations mutuelles de cooccurrence des mots sur des fenêtres de taille fixe. Nous choisissons pour ce faire les fenêtres de taille 10 (window10).

En outre, nous observons dans les définitions la présence optionnelle d'exemples tels que *constitution d'un dossier, d'un gouvernement* dans la définition du sens I de *constitution* ou bien *constitution d'une dot, d'une rente* dans la définition du sens III.1. Ces exemples ne sont pas identifiables en tant que tels dans les descriptions car aucune structure ne les distingue de la définition proprement dite. Même le fait qu'ils soient écrits sur une autre ligne n'est pas exploitable puisqu'ils apparaissent respectivement à la ligne 2 et 3.

Nous proposons néanmoins d'utiliser cette information en procédant à un traitement spécifique distinct lorsque la définition contenant le mot dont on cherche les sens est présent dans la définition (comme dans les définitions I et III.1) et est en relation syntaxique avec un autre mot de la définition (e.g. *dot* dans la définition I). On ajoute alors au score des facteurs issus de l'information mutuelle de cooccurrence syntaxique des mots du cluster et du mot en relation syntaxique (e.g. ici, on calculerait l'information mutuelle entre *mot_du_cluster* et *dot* dans la matrice gouvernée par la relation *COMPDUNOM*).

Ainsi, nous avons procédé suivant le pseudocode décrit à l'algorithme 1 où l'ensemble $I_d$ correspond à l'ensemble des lemmes des mots plein de la définition $d$, $J_c$ correspond à l'ensemble des lemmes appartenant au cluster $c$, et $\mathcal{RELS}_d = \{(l_i, rel_k)\}$ correspond à l'ensemble des paires dont le premier élément est le lemme de la définition $d$ lorsqu'il est en relation syntaxique avec le mot ambigu, et le deuxième élément est ladite relation syntaxique.

#### 4.2.1.3 Validation croisée

Un certain nombre de paramètres étant à fixer à la fois lors de la phase d'induction de sens et lors de la phase de désambiguïsation elle même, nous souhaitons procéder à un apprentissage des meilleurs paramètres sur un corpus d'entraînement. Nous procédons à une validation croisée de nos résultats sur 10 échantillons (*10-fold cross validation*) en s'assurant que chaque *fold* contient autant d'exemples que les autres pour chaque mot donné (à un près).

Pour l'algorithme MultiNN, nous avons fait varier cinq paramètres : le quota de votants (cf. section 3.2.3.1), le seuil de Faiblesse, le seuil de Jointure, le seuil de Noyau, le seuil de Bruit et enfin le ratio

**Algorithm 1** Mapping des clusters de sens induits aux sens Romanseval

**Require:** motsAmbigus, definitionsRomanseval, clustersWSI
**Ensure:** mapping
  **for all** m in motAmbigus **do**
    **for all** c in clustersWSI[m] **do**
      **for all** d in definitionsRomanseval[m] **do**

$$score_c[d] = \sum_{l_j \in J_c} \sum_{l_i \in I_d} association(l_i, l_j)$$

avec :

$$association(l_i, l_j) = \begin{cases} \alpha_i.PMI_{rel_k}(l_i, l_j)) & si(l_i, rel_k) \in \mathcal{RELS}_d \\ \alpha_i.PMI_{window10}(l_i, l_j)) & sinon \end{cases}$$

et :

$$\alpha_i = \begin{cases} \dfrac{2}{|I_d|} & si(l_i, rel_k) \in \mathcal{RELS}_d \\ \dfrac{1 - \frac{2.|\mathcal{RELS}_d|}{n}}{|I_d| - |\mathcal{RELS}_d|} & sinon \end{cases}$$

si bien que $\sum_{i=1}^{|I_d|} \alpha_i = 1$

      **end for**
      $mapping[c] = argmax_{d \in definitions[m]}(score_c[d])$
    **end for**
  **end for**
  **return** mapping

d'adaptation de la distance du voisinage, ainsi que les paramètres $k$ et $modeCV$ de l'algorithme de désambiguïsation lui-même.

Pour l'algorithme MultiHyperLex nous avons étudié l'influence des cinq paramètres suivants : le ratio de votants, la taille minimale des clusters pour qu'ils soient validés, la taille maximale au-delà de laquelle une ré-assignation est effectuée, le pas d'augmentation de ce seuil, et enfin le mode de choix du noyau suivant, ainsi que les paramètres de l'algorithme de désambiguïsation. Les modes de choix du noyau suivant appartiennent à deux méthodes distinctes. Tous les éléments non encore attribués à un cluster sont candidats. Dans la méthode *closestElement*, on choisit l'élément le plus proche de tous les autres : soit en choisissant le minimum des moyennes de similarités aux autres éléments le plus élevé parmi les différents espaces ($minS$), *i.e.* tous les autres candidats ont au moins un espace dans lequel leur moyenne des similarités est plus faible, soit en choisissant le maximum le plus élevé ($maxS$), *i.e.* il n'existe pas de candidat et d'espace pour lesquels cette moyenne des similarités soit plus élevée, soit en choisissant la moyenne la plus élevée ($moyS$), *i.e.* globalement cette moyenne des similarités est bonne parmi les différents espaces. La méthode *bestHub* exploite également ces trois façons de choisir le noyau mais cette fois-ci en optimisant la variance (cf. 3.2.3.2), donnant ainsi les trois variantes $minV$, $maxV$, $moyV$.

| Quota | Seuil Bruit | Seuil Noyau | Seuil Faiblesse | Seuil Jointure | k | modeCV |
|---|---|---|---|---|---|---|
| 0,2 | 0,4 | 0,75 | 0 | 0.10 | 10 | D |

TABLE 4.2 – Meilleure configuration pour la désambiguïsation à l'aide des sens issus de MultiNN

| Quota | Taille Minimale | Taille Maximale | Pas | Ratio | Choix Noyau | k | modeCV |
|---|---|---|---|---|---|---|---|
| 0,6 | 3 | 15 | 5 | 2 | minS | 95 | D |

TABLE 4.3 – Meilleure configuration pour la désambiguïsation à l'aide des sens issus de MultiHyperLex

Pour les deux algorithmes, dans les dix cas c'est la même configuration qui a donné lieu à la meilleure F-mesure *fine-grained* sur le corpus d'entraînement. Cette configuration est donnée dans les tableaux 4.2 et 4.3.

Les résultats présentés en figure 4.2 correspondent à la moyenne des mesures obtenues sur ces 10 évaluations en utilisant les paramètres ayant donné le plus de fois la meilleure précision sur les échantillons d'entraînement. Ces données sont mises en parallèle avec les résultats produits par les deux baselines proposées par Romanseval.

Romanseval proposait une première baseline (notée *Baseline*) correspondant au premier sens fourni par le dictionnaire, ainsi qu'une seconde baseline (notée *Cheap*) correspondant à une variante simplifiée de l'algorithme de Lesk consistant à attribuer la définition contenant le plus de mots en commun avec le contexte du mot ambigu.

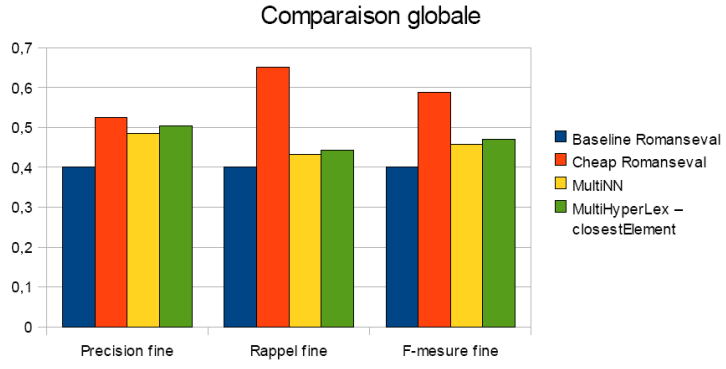



FIGURE 4.2 – Comparaison globale des systèmes

#### 4.2.1.4 Analyse des résultats

Les deux algorithmes nous donnent une F-mesure inférieure à la deuxième *baseline* proposée dans Romanseval (cf. [Segond 2000]). Il faut noter que lors de la campagne cette seconde baseline est arrivée en $3^{ème}$ position sur 7 pour la désambiguïsation des noms.

Nous observons maintenant le détail de la F-mesure obtenue pour chacun des mots. Le détail est rapporté à la figure 4.3. Nous allons analyser les traitements pour lesquels les mots ont obtenus une très faible F-mesure (inférieure à 20%).

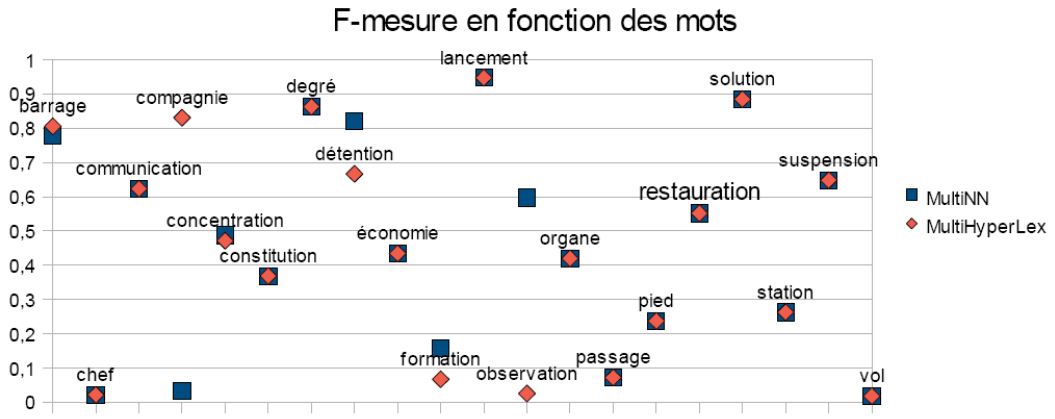



FIGURE 4.3 – Une désambiguïsation hétérogène

**chef** Pour le mot *chef*, le tableau 4.4 présente d'une part la définition de l'annotation utilisée majoritairement (pour 89% des instances ambiguës) par l'annotation manuelle, et d'autre part les clusters ayant été utilisé par nos algorithmes pour 100% des instances du mot *chef*. Les deux algorithmes annotent toutes les instances du mot *chef* avec un cluster paraissant bien représenter le sens annoté manuellement dans 89% des instances du mot *chef*. Pourtant la précision rapporté est presque nulle. Le problème ne vient donc ni de la définition des sens, ni de la désambiguïsation, mais plutôt de la phase de *mapping*.

En effet, le cluster en question est mis en correspondance avec le sens *Petit Larousse* dont la définition est : *de son chef , de son propre chef : de sa propre autorité*. Cette définition ne comporte que très peu de mots pleins. Chaque proximité ou distance au mot de la définition compte donc d'autant plus pour l'algorithme de mapping. La distance à la définition n'est donc pas bruitée par d'autres mots un peu plus hors-contexte. Le mot *autorité* apparaît souvent dans les contextes des éléments de notre cluster (dans le corpus de constitution de l'espace sémantique). Ainsi, le mot *autorité* donne un score très élevé à cette définition et les clusters dont nous parlons sont mis en correspondance avec ce sens, induisant ainsi une erreur de désambiguïsation présente pour 89 %

| | |
|---|---|
| Annotation manuelle | personne qui commande, qui exerce une autorité, une direction, une influence déterminante.<br>chef de famille.<br>chef d'entreprise. |
| Annotation MultiNN | coach , gamin, cheikh, lieutenant , porte-parole , vizir , artiste , centre , clan, commandant, conseil, contraire, dernier, entreprise, équipe, famille, femme, fille, frère, gens, gouvernement, grand, groupe, guide, homme, hôtel, jeune, joueur, juge, loi, maire, maître, membre, mère, ministre, modèle, mouvement, musée, musicien, opposition,(...) |
| Annotation MultiHyperLex | prince, reine, prophète, roi, juge, témoin, président, parti, ministre, organisation, gouvernement, jeune, gens, mouvement, entreprise, ville, équipe, vie, pays, membre, femme, personne |

TABLE 4.4 – Annotations majoritaires pour le mot *chef*

des instances de *chef*. On peut clairement avancer que dans le cas du mot *chef*, la principale source d'erreur provient de l'étape de *mapping*.

**compagnie** Une fois de plus, pour le mot *compagnie*, 90% des annotations manuelles concernent un seul et même sens. La définition de ce sens est la suivante : *société commerciale assurant un service public. compagnie d'assurances. - et compagnie, s'ajoute à une raison sociale après l'énumération des associés nommés (abréviation : et cie).*

Cette fois-ci l'algorithme MultiHyperLex dispose d'un cluster parfaitement adapté à la définition et mis en correspondance sans problème. Ce cluster est constitué des éléments :

*assureur, équipementier, automobiliste, financier, négociant, armateur, multinationale, banquier, mairie, brigade, mutuelle, traducteur, courtier, transporteur, filiale, commerçant, firme, syndicat, commission, gouvernement, institution, conseil, société, secteur.*

L'algorithme MultiNN dispose d'un cluster équivalent :

*assureur, armateur, arthropode, automobiliste, garagiste, magnat, rallye, banquier, commerçant, courtier, multinationale, mutuelle, négociant, transporteur.*

Celui-ci est correctement mis en correspondance avec les sens *Petit Larousse* mais l'algorithme de désambiguïsation ne choisit pas ce cluster pour l'annotation. Le choix se porte un autre cluster contenant également des éléments en rapport avec une *compagnie commerciale*, mais dont une grande partie des éléments correspond à une *compagnie militaire*. Le mapping sur ce cluster resté ambigu a produit un lien vers le sens *Petit Larousse* de *compagnie militaire* et toutes les annotations concernées sont donc erronées. Ici c'est la qualité des clusters issus de l'algorithme MultiNN qui a posé un problème.

**formation** Une nouvelle fois, 88% des annotations manuelles concernent un seul et même sens défini de la façon suivante : *action de former quelqu'un intellectuellement ou moralement ; instruction, éducation. - formation permanente ou continue : formation professionnelle destinée aux salariés*

*des entreprises. - formation professionnelle : ensemble des mesures adoptées pour la formation des travailleurs, prises en charge par l'État et les employeurs.*

Nos clusters choisis pour l'annotation sont mis en correspondance avec un sens très proche dont la définition est la suivante :

*ensemble des connaissances dans un domaine déterminé ; culture. formation littéraire. il n'a aucune formation.*

Dans le cas de ce mot, c'est la granularité de distinction qui est trop fine pour notre désambiguïsateur. Même si la mesure utilisée pour cette analyse n'est pas la même (kappa-mesure), on peut voir sur la figure 4.4 que ce mot a également posé des difficultés aux participants de la campagne Romanseval puisque c'est le cinquième mot le plus mal désambiguïsé.

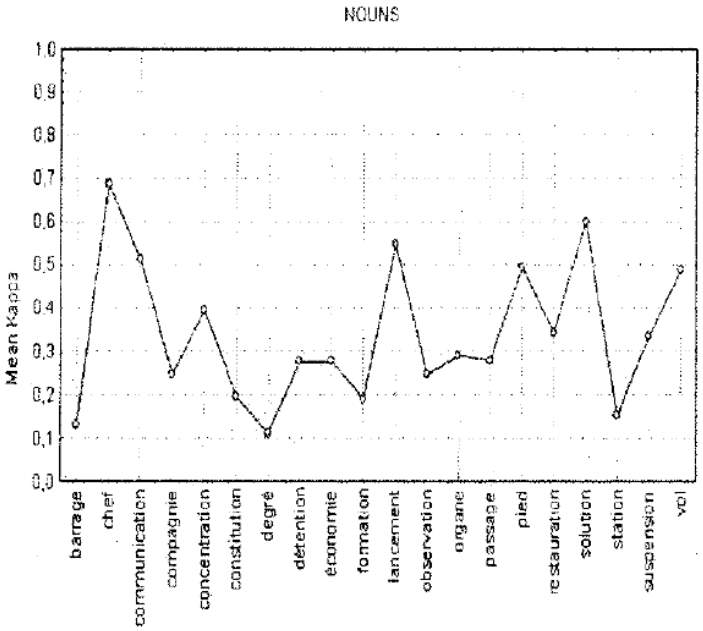



FIGURE 4.4 – Moyenne des mesures Kappa des systèmes ayant participé à Romanseval en fonction des mots ambigus. [Segond 2000]

**observation** Une fois de plus, le mauvais mapping d'un cluster correspondant au sens majoritaire annoté manuellement (sur 96% des instances) est la source du problème pour les sens issus de MultiHyperLex.

Le cluster constitué des éléments :

*sous-ministre, bibliographie, allégation, chasseur, représentation, échange, idée, examen, intervention, donnée, réponse, résultat, question, lecture, disposition, conception, connaissance, mouvement, réflexion, mesure, expérience, recherche, approche, information, collaboration, critère, évaluation, étude, évolution, définition, ordre* ,

est mis en correspondance avec le sens de définition :

*militaire. surveillance systématique de l'ennemi en vue d'obtenir des renseignements* . au lieu du sens de définition *compte rendu, ensemble de remarques, de réflexions de quelqu'un qui a observé , étudié quelque chose. consigner ses observations sur un registre.*

En revanche, les sens produits par MultiNN et utilisés pour l'annotation sont correctement mis en correspondance. L'étape de mapping est à nouveau le facteur bloquant d'une bonne désambiguïsation.

**passage** L'annotation dont la définition est :

*action, fait de passer. le passage des hirondelles. le passage du rire aux larmes. - droit droit de passage : droit de passer sur la propriété d'autrui. (la servitude de passage est légale en cas d'enclave)*

est acceptée par les annotateurs humains pour 100% des instances du mot *passage*. Or cette définition contient des mots comme *servitude, légale* ou *enclave*, que l'on associe assez rarement avec le mot *passage* dans un contexte général. Cela rend d'une part le mapping avec ce sens très difficile. D'autre part l'examen manuel des clusters automatiques choisis comme annotation montre des clusters très gros et très ambigus. Nous pensons que cela est dû au choix des mots à clusteriser qui, pour le mot *passage* sont très généraux et souvent polysémiques.

**vol** L'analyse de ce dernier mot ne fait pas exception. Le tableau 4.5 nous montre la difficulté de mettre en correspondance les sens de granularité la plus fine. Les clusters issus de l'induction automatique sont mis en correspondance avec les sens *Petit Larousse*. On observe que pour la granularité de haut niveau la mise en correspondance est correcte pour les deux clusters présentés, par contre la mise en correspondance devient incorrecte lorsque l'on passe au niveau le plus fin.

L'analyse de ces cas d'erreurs les plus systématiques nous montre que très souvent, c'est le mapping erroné qui a provoqué l'échec de la désambiguïsation. Nous n'avons pas procédé à un mapping manuel car il ne peut être effectué en conditions réelles pour tous les mots du vocabulaire ni pour procéder à l'optimisation des meilleurs paramètres. L'évaluation par la V-mesure de la sous-section suivante nous permettra de nous abstraire de l'implication de ce mapping dans l'évaluation.

On peut également remarquer que l'algorithme MultiNN préfère utiliser un $k$ de petite valeur dans l'algorithme de désambiguïsation tandis que l'algorithme MultiHyperLex a tendance à utiliser un

| | | |
|---|---|---|
| Clusters issus de MultiHyperLex choisis lors de l'annotation automatique | biréacteur, fusée, fret, voilier, navette, missile, voyagiste, planeur, cargo, paquebot, bombardier, avion | fourberie, illégalité, gourmandise, acte, duplicité |
| Sens *Petit Larousse* obtenus par le mapping automatique | groupe d'oiseaux qui volent ensemble. | fait de prendre plus que ce qui est dû, de vendre à un prix excessif. |
| Sens choisi par les annotateurs Romanseval | déplacement actif dans l'air des oiseaux, des insectes, etc., au moyen de surfaces latérales battantes (ailes).<br>- vol ramé, dans lequel les ailes s'appuient sur l'air comme les rames sur l'eau. - vol plané, dans lequel les ailes glissent sur l'air.<br>- vol à voile, qui utilise la puissance du vent et ses courants ascendants.<br>- vol bourdonnant, à battements très rapides, permettant le maintien en un point fixe (insectes, colibri).<br>ou<br>déplacement dans l'air d'un aéronef ou dans l'espace d'un engin spatial ; l'engin lui-même.<br>- descendre en vol plané, moteur arrêté.<br>- vol à voile : mode de déplacement d'un planeur utilisant les courants aériens.<br>- terme du vocabulaire sportif.<br>vol libre, pratiqué avec une aile libre. | action de voler , de dérober ce qui appartient à autrui. |

TABLE 4.5 – Mise en correspondance des sens du mot *vol*

$k$ très élevé. Cette différence peut s'expliquer par le fait que les clusters produits par l'algorithme MultiNN sont en moyenne beaucoup plus gros que ceux produits par l'algorithme MultiHyperLex. Ainsi l'algorithme de désambiguïsation utilisant les sens de MultiNN n'a pas besoin d'aller chercher très loin dans la sphère de voisinage pour trouver des éléments appartenant à nos classes de désambiguïsation. Ceci n'est pas le cas pour l'algorithme utilisant les clusters issus de MultiHyperLex, l'algorithme préfère donc considérer une sphère de voisinage plus grande.

### 4.2.2 Évaluation par la V-mesure

Dans le but d'évaluer notre système de désambiguïsation sans prendre en compte la phase de *mapping* et également afin d'obtenir une idée de la pertinence de notre système par rapport à des systèmes état de l'art exploitant aussi des sens de mots induits, nous comparons nos résultats à ceux de la tâche *Word Sense Induction* de la campagne SemEval 2010 ([Manandhar *et al.* 2010]). Bien que les répertoires de sens et les données utilisées ne soient pas les mêmes, les résultats de la campagne SemEval permettent de donner un ordre de grandeur des mesures atteintes par les systèmes état de l'art pour des mesures spécifiques aux clusters de sens.

#### 4.2.2.1 Protocole expérimental

Reprenant une partie du protocole d'évaluation de [Manandhar *et al.* 2010], nous calculons la V-mesure (définie par [Rosenberg & Hirschberg 2007]) des clusters produits après désambiguïsation. En effet, une telle mesure quantifie la qualité des clusters par rapport à un ensemble de classes données constituant la vérité terrain. La vérité-terrain nous est ici fournie par le corpus de test de la campagne Romanseval où chaque instance ambiguë est annotée par son sens *Petit Larousse*.

On distinguera les clusters primaires constitués par nos sens induits à la première étape, des clusters secondaires, chaque cluster secondaire étant l'ensemble des instances ayant été annotées par le même cluster par notre système. Ainsi, ce sont les clusters secondaires qui seront comparés à la vérité terrain pour fournir les mesures souhaitées.

Nous nous intéressons aux mesures d'homogénéité : est-ce que les clusters ne mélangent pas trop les sens ? ; et de complétude : est-ce que les classes de sens ne sont pas présentes dans trop de clusters différents ? Ces mesures fondées sur l'entropie sont données par les formules suivantes. Soient $K = \{k_i\}$ l'ensemble des clusters secondaires et $C = \{C_j\}$ l'ensemble des classes. On a alors :

$$homogénéité = \begin{cases} 1 & si\ H(C) = 0 \\ 1 - \dfrac{H(C|K)}{H(C)} & sinon \end{cases}$$

$$complétude = \begin{cases} 1 & si\ H(K) = 0 \\ 1 - \dfrac{H(K|C)}{H(K)} & sinon \end{cases}$$

$$V - mesure = 2.\frac{homogénéité * complétude}{homogénéité + complétude}$$

avec :

$$H(C|K) = -\sum_{k=1}^{|K|}\sum_{c=1}^{|C|}\frac{a_{ck}}{N}log\frac{a_{ck}}{\sum_{c=1}^{|C|}a_{ck}}$$

$$H(C) = -\sum_{c=1}^{|C|}\frac{\sum_{k=1}^{|K|}a_{ck}}{N}log\frac{\sum_{k=1}^{|K|}a_{ck}}{N}$$

$$H(K|C) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{a_{ck}}{N}log\frac{a_{ck}}{\sum_{k=1}^{|K|}a_{ck}}$$

$$H(K) = -\sum_{k=1}^{|K|}\frac{\sum_{c=1}^{|C|}a_{ck}}{N}log\frac{\sum_{c=1}^{|C|}a_{ck}}{N}$$

et $N$ le nombre d'instances ambiguës.

Dans le cadre de Romanseval, toute annotation faite par un annotateur humain est considérée comme correcte si bien qu'il peut y avoir plusieurs classes attribuées à chaque instance. Ainsi nous proposons un ajustement permettant d'approximer une V-mesure dite *généreuse*. Pour chaque mot on calcule d'une part l'homogénéité et la complétude en choisissant parmi les classes données en annotation les classes maximisant l'homogénéité (maximum local), d'autre part en choisissant les classes maximisant la complétude (maximum local).

Nous choisissons également de borner ces mesures en calculant la V-mesure dite *stricte* obtenue à partir des annotations strictes, c'est-à-dire quand on considère comme vérité l'annotation la plus fréquemment donnée par un annotateur. En cas d'égalité entre plusieurs annotations, la même procédure d'ajustement que pour la V-mesure généreuse est appliquée.

L'homogénéité et la complétude sont calculées distinctement pour chaque mot puis la moyenne est faite sur l'ensemble des mots pour donner les bornes finales (homogénéité stricte, homogénéité généreuse, complétude stricte, complétude généreuse), et finalement calculer les bornes de la V-mesure.

#### 4.2.2.2 Résultats

On calcule ces mesures sur les résultats obtenus sur l'intégralité du corpus en utilisant les deux configurations retenues lors de notre évaluation Romanseval. Les résultats sont présentés dans le tableau 4.6.

Ces résultats montrent la qualité de notre désambiguïsation si l'on s'abstrait de la phase de mapping. Bien que ce ne soient pas les mêmes données ni le même référentiel de sens, nous pouvons

| Mesures | Homogénéité | | Complétude | | V-mesure | |
|---|---|---|---|---|---|---|
| Bornes | min | max | min | max | min | max |
| MultiNN | 45,2 | 64,4 | 16,2 | 40,3 | 23,9 | 49,6 |
| MultiHyperLex | 39,0 | 60,2 | 14,8 | 38,9 | 21,5 | 47,3 |

TABLE 4.6 – Bornes de la V-mesure

mettre en perspective nos résultats par rapport à ceux de la tâche d'induction de sens de SemEval 2010 ([Manandhar *et al.* 2010]). Le système ayant obtenu le meilleur résultat sur les noms a obtenu une V-mesure de 20,6, ce qui est légèrement inférieur à notre borne minimale. Cela est loin de signifier que notre algorithme est meilleur que ceux présentés lors de la campagne d'évaluation, mais cela montre que malgré les résultats présentés comme inférieurs à ceux de la deuxième baseline Romanseval, nos systèmes produisent une désambiguïsation très pertinente lorsque l'on n'a pas besoin de recourir à une étape de mise en correspondance des clusters de sens à des sens produits manuellement.

### 4.2.3 Conclusions

L'algorithme de désambiguïsation que nous proposons intègre directement son propre référentiel de sens, issu d'une analyse automatique.

Nous avons pu voir au travers de l'analyse des erreurs, les différentes difficultés rencontrées lors de la mise en correspondance de clusters de sens avec un référentiel de sens constitué manuellement et sans structure particulière.

Enfin, la comparaison de la V-mesure obtenue sur nos résultats avec celles obtenues par les systèmes anglophones semble indiquer que pour une application ne nécessitant pas de *mapping*, notre algorithme produit des résultats au moins équivalents à ceux de l'état de l'art des systèmes utilisant des sens induits.

Ce dernier résultat nous encourage fortement à intégrer notre algorithme dans un cadre plus applicatif afin de déterminer s'il peut effectivement aider à améliorer les systèmes de recherche d'information.

## 4.3 Intérêt de la désambiguïsation en Recherche d'Information

Nous revenons donc à notre problématique initiale de recherche d'information. Dans le but d'améliorer les systèmes de recherche, nous proposons deux mises en œuvre de désambiguïsation relatives d'une part aux moteurs de recherche classiques et d'autre part aux systèmes de réponses aux questions. Dans un cas comme dans l'autre, cette intégration ne requiert pas de *mapping* à un

référentiel de sens constitué manuellement.

#### 4.3.0.1 Indexation par sens

La première intégration consiste à indexer les sens des mots relatifs à nos clusters lors de l'indexation. Lorsqu'un utilisateur entre une requête, si c'est une requête ne contenant qu'un unique terme, le moteur renvoie dans un premier temps les résultats traditionnels liés à l'indexation de mots. Le système propose alors à l'utilisateur parmi les divers raffinements la possibilité de choisir le ou les sens auxquels sa requête fait référence. Ces différents sens sont proposés sous forme de sac de mots correspondant à nos clusters. Si c'est une requête comportant plusieurs termes et éventuellement des relations syntaxiques entre ces termes, le système tente de désambiguïser la requête, renvoie les résultats correspondant au sens reconnu s'il en existe (sinon il renvoie tous les résultats), mais propose également à l'utilisateur de modifier la reconnaissance du sens s'il le souhaite, par la même interface que lorsqu'il s'agit d'une requête mono-terme. Nous voyons dans cette interaction un moyen pour l'utilisateur de modifier la granularité de distinction de sens et également de pallier le problème de précision lorsque le contexte de la requête est trop pauvre pour que l'analyse soit correcte.

Il serait également intéressant d'indexer et de proposer à l'utilisateur le choix de la catégorie syntaxique des mots de la requête. Ces catégories syntaxiques sont parfois fortement discriminantes et génèrent moins d'erreurs d'analyse dans la désambiguïsation.

#### 4.3.0.2 Systèmes de Q/R

Le deuxième cadre dans lequel la désambiguïsation peut s'avérer intéressante est celui de la recherche de réponses aux questions posées en langage naturel. Les mots ambigus sont alors très souvent dans un contexte discriminant. En indexant également le sens des mots, une désambiguïsation des mots ambigus de la question permettrait de donner un poids plus fort aux extraits contenant les mots dont le sens indexé correspond au sens analysé dans la question.

À la question *Combien pèse une hotte de mineur ?*[2], le mot *hotte* pourrait être reconnu comme un contenant que l'on porte sur le dos et ainsi permettre d'éliminer les occurrences de *hotte* en tant que *hotte de cuisine* (très présentes parmi les premières réponses de Google). Il en va de même pour le mot *mineur* très présent en tant qu'adjectif opposé à *majeur*.

Qu'en est-il de l'extension des mots de la requête ? Dans une situation idéale, on pourrait penser que la désambiguïsation combinée à un mapping sur les sens de Jaws permettrait d'étendre la requête aux synonymes restreints en fonction des sens employés. Cette stratégie n'a pas été testée mais nous pensons qu'elle est encore très précoce du fait de la perte de qualité liée à la succession d'étapes. En effet on peut s'interroger sur la pertinence des traitements à tous les niveaux : le sens correct

2. question issue de l'évaluation Quaero Q/R

est-il présent dans les clusters ?, la désambiguïsation des mots de la question est-elle correctement effectuée ?, le sens correspondant existe-t-il dans Jaws ?, existait-il dans WordNet ?, le mapping du cluster concerné est-il fait correctement avec les sens de Jaws ?, quelle est la qualité des synonymes présents dans Jaws pour ce sens ? En pratique, l'implémentation de cette stratégie n'est pas encore réalisable mais nous pensons que l'amélioration de chacune de ces étapes mènerait à une amélioration des systèmes de Q/R français.

#### 4.3.0.3 Bilan

Bien que nous n'ayons pu mettre en place de tels prototypes, nous pensons que le système de désambiguïsation que nous proposons pourrait permettre un enrichissement des systèmes de recherche d'information dans un cadre d'interaction avec l'utilisateur. Dans un tel cadre, l'utilisateur se sert de l'information que nous lui fournissons mais peut aussi se dégager de l'analyse automatique de la requête si celle-ci fait défaut. L'implémentation d'un tel prototype permettrait de valider cette hypothèse.

Pour un usage plus automatique, la présence d'un contexte riche dans les systèmes de Q/R permettra une désambiguïsation plus robuste, utile pour restreindre l'extraction des paragraphes à ceux concernant les sens des mots employés dans la question. En revanche, en ce qui concerne l'enrichissement de la requête, les ressources françaises n'étant pas encore suffisamment mûres pour une extension de requête par une désambiguïsation explicite, il faudra penser à d'autres stratégies (cf. voir section 2.5.2 de l'état de l'art).

## 4.4 Conclusions

Le système de désambiguïsation lexicale que nous proposons est un des rares systèmes de désambiguïsation français exploitant directement des clusters de sens. En comparaison aux systèmes anglais de désambiguïsation spécifique aux sens induits évalués lors de la campagne SemEval 2010, notre système donne des résultats très encourageants.

Bien qu'il exploite en interne une technique de classification supervisée (k-NN multi-représenté), notre système complet de désambiguïsation présente l'avantage de ne pas nécessiter de données désambiguïsées en entrée. Il est également intéressant de noter que les seules informations que notre système utilise sont issues d'espaces distributionnels construits automatiquement à partir d'une analyse syntaxique. Il est donc en théorie transposable à toute langue disposant d'un analyseur syntaxique et de corpus de texte suffisamment grands.

Nos résultats montrent que les sens des mots dans un texte peuvent être discriminés par une telle technique, cependant l'expérience a également montré que ceux-ci sont difficilement liables à des

définitions. Ainsi nos sens de mots sont eux-mêmes représentés par des sacs de mots et restent assez imprécis pour quelqu'un qui ne connaît pas préalablement les différents sens existants.

Supposons que l'on ignore les sens du mot *pêche* et que l'on trouve les trois sacs de mots {*poire*, *abricot*, *nectarine*, (...)}, {*chasse*, *culture*, *commerce*, (...)} et {*écotourisme*, *farniente*, *équitation*, (...)}. La connaissance des autres mots nous fournira l'information suffisante pour inférer que la *pêche* peut être un fruit, une activité économique ou de loisir. Cependant, il reste complètement impossible de faire le lien avec le *poisson*. Cela est pourtant le cas lorsque l'on inclut également des cooccurrents dans les mots que l'on clusterise, mais à nouveau, il nous manque une information de compositionnalité : on ne peut pas savoir ce qu'il faut faire avec le *poisson*.

Ainsi, si l'on considère une langue que l'on connaît relativement bien sans pour autant connaître tous les sens des mots, un dictionnaire nous permettra de comprendre le sens inconnu d'un mot dans une phrase, tandis qu'un sac de mots issus des plus proches voisins nous fournira une information d'analogie mais ne définira pas le mot inconnu.

Ce paradigme n'est donc pas adapté pour un humain qui cherche un sens. En revanche, comme nous l'évoquions dans la dernière section de ce chapitre, cela n'est plus nécessairement un désavantage lorsque ce système devient une brique interne d'une application plus complexe. En effet, dans le cadre d'une interface de recherche de documents, l'utilisateur possède la connaissance préalable des définitions des mots (à moins qu'il ne cherche justement une définition) et saura quel cluster correspond au sens du mot qu'il cherche.

Par ailleurs, dans le cas des systèmes de réponses à des questions posées en langue naturelle, le système effectue seul la désambiguïsation des mots à la fois dans la question et dans les passages comportant potentiellement la réponse. Le référentiel de sens étant le même, le système n'a pas besoin de connaître de définitions des sens pour faire correspondre un sens reconnu dans une question avec le même sens reconnu dans un passage candidat. Cette approche par clusters de mots constitués automatiquement n'est donc plus un inconvénient mais devient même un avantage, particulièrement pour discriminer les ambiguïtés sur des néologismes (la *toile* pour les contenus d'Internet), ou sur des mots plus éphémères comme les usages familiers (*c'est frais* pour signifier un enthousiasme particulier), les marques (*Orange*), les personnalités connues (*Paris Hilton*), etc.

Au final, nous avons produit deux ressources françaises à large échelle. La première étant construite sur la structure de WordNet, nous espérons qu'elle pourra être couramment utilisée par la communauté pour les tâches qui utilisent déjà le WordNet de Princeton pour le traitement de l'anglais. La seconde ressource générée est un ensemble de clusters de mots représentant des sens. Cette ressource nous a permit de bâtir notre propre système de désambiguïsation, indépendant de tout référentiel de sens manuel. Ce système de désambiguïsation non supervisé donne des résultats très encourageants. Nous pensons intégrer ce système résultant de notre étude dans des applications de recherche d'information chez Exalead.

La désambiguïsation lexicale porte sur la distinction du sens des mots indépendamment les uns les autres. Elle permet d'effectuer un certain type d'inférence logique notamment en ce qui concerne les termes polysémiques qui sont en relation sémantiques avec d'autres termes. Ceci est notamment le cas lorsque le terme polysémique est lui-même un hyperonyme et que l'on souhaite étendre ses propriétés à ses hyponymes ou à ses instances, c'est-à-dire aux concepts qu'il englobe. Par exemple, la désambiguïsation du mot *console* dans la phrase *Une console de jeu est un cadeau très populaire chez les adolescents.* permet d'inférer que *La XBox 360 doit être un cadeau très populaire chez les adolescents.* tandis qu'une *console en acajou* ou un *terminal X* ne le sont pas nécessairement.

Cette désambiguïsation lexicale n'est cependant pas suffisante pour permettre une totale interprétation du texte. Lorsque l'humain entend une phrase et l'interprète, il fait également le lien entre les différents mots de la phrase et détermine inconsciemment pourquoi ces mots sont utilisés les uns avec les autres, autrement dit à quoi un mot sert à un autre, quelle est la signification de ce qui les relie. C'est cette dimension supplémentaire qui permet d'effectuer des inférences logiques plus complexes. Nous abordons ainsi cette seconde thématique dans la troisième partie de notre manuscrit.

# Troisième partie

# Annotation en rôles sémantiques

# Chapitre 5

# Création de ressources pour l'annotation sémantique de rôles

Nous avons vu qu'une approche de l'analyse sémantique peut aussi consister à annoter un texte en rôles sémantiques (*Semantic Role Labeling - SRL*). Plusieurs ressources sémantiques ont été construites pour décrire des ensembles de situations standard associées à des ensembles de rôles que les syntagmes du texte peuvent remplir et nous les avons répertoriées et décrites à la section 2.2.1.3 de notre état de l'art. L'annotation sémantique de rôles consiste à attribuer des rôles sémantiques décrits dans ces ressources aux syntagmes des phrases analysées.

Rappelons que dans FrameNet ([Baker *et al.* 1998]), chaque frame est définie par rapport à un ensemble de rôles qui lui est spécifique. Ces rôles sont appelés *frame elements*.

FrameNet répertorie aussi des prédicats (verbes, noms, adjectifs et même adverbes) appelés unités lexicales (*Lexical Units - LU*) déclenchant ces situations, autrement dit les prédicats régissant ces ensembles de rôles.

Par exemple la figure 5.1 représente une partie de la frame */Ingestion/*. Cette frame comprend deux rôles principaux (*Core*) : *Ingestor* et *Ingestibles*[1], ainsi que des rôles secondaires (*Non-Core*) tels que *Degree*, *Duration*[2], *Instrument*.

La ressource contient aussi une liste d'unités lexicales déclenchant cette frame, on a par exemple : *devour.v*, *breakfast.v*, *consume.v*, *dine.v*, *drink.v*, *sip.v*, *sip.n*[3], etc.

Nous nous intéressons plus particulièrement à la ressource FrameNet car la généralisation de différentes unités lexicales à une même frame permet de rapprocher deux formulations différentes

1. Ingestor : *celui qui ingère*, Ingestibles : *ce qui est ingéré*
2. Degree : *degré*, Duration : *durée*
3. devour : *dévorer*, breakfast.v : *prendre son petit déjeûner*, consume : *consommer*, dine : *dîner*, drink :*boire*, sip.v :*siroter*, sip.n :*petite gorgée*

## Ingestion

Definition:
*An* Ingestor *consumes food or drink* Ingestibles*, which entails putting the* Ingestibles *in the mouth for delivery to the digestive system. This may include the use of an* Instrument*. Sentences that describe the provision of food to others are NOT included in this frame.*

**FEs:**

**Core:**

- Ingestibles The Ingestibles are the entities that are being consumed by the Ingestor.
  *The wolves* ***DEVOURED*** *the carcass completely.*
- Ingestor The Ingestor is the person eating or drinking.
  *The wolves* ***DEVOURED*** *the carcass completely.*

**Non-Core:**

- Degree The extent to which the Ingestibles are consumed by the Ingestor.
  *The wolves* ***DEVOURED*** *the carcass completely.*
- Instrument The Instrument with which an intentional act is performed.

FIGURE 5.1 – Frame *Ingestion* et ses différents *Frame Elements (FEs)*

d'un même fait, ce qui peut être très utile dans des systèmes de Q/R. Cela permet par exemple pour répondre à la question *Qui a inventé le cinéma ?* d'extraire les phrases *Tel qu'on le connaît aujourd'hui, le cinéma* ***a été inventé*** *par les frères Georges et Louis Lumière.* et *Louis, avec son frère Auguste et sous les conseils de leur père Antoine,* ***ont conçu*** *ce cinématographe qui est monté sur une manivelle.*, sans avoir à générer explicitement de paraphrases.

En effet, a présence du verbe *inventer* et du verbe *concevoir* dans la même frame permettrait une analyse similaire de ces deux phrases. En outre, la détection du rôle d'*Inventeur* à la fois dans la question et dans les deux phrases extraites permettrait une extraction fiable de la réponse. Ce rôle étant réalisé par le mot interrogatif *qui* typant la question, la détection des arguments *par les frères Georges et Louis Lumière* et *Louis, avec son frère Auguste et sous les conseils de leur père Antoine* identifient clairement les réponses dans les deux phrases données.

Nous avons vu dans notre état de l'art (2.3.1.2) que les ressources francophones pour le SRL sont rares. Nous ferons référence dans ce chapitre à la seule ressource francophone de type FrameNet ([Padó & Pitel 2007]), et la nommerons FrameNet.Fr.

[Padó & Pitel 2007] utilisent une projection par l'usage de corpus parallèles bilingues alignés. Dans l'étude précédente de [Padó & Lapata 2005] l'approche par projection appliquée sur la langue allemande est également comparée avec une autre méthode utilisant un dictionnaire bilingue. Les résultats de cette dernière sont moins bons. Nous croyons néanmoins en la possibilité d'une approche par dictionnaires bilingues et proposons dans ces travaux une nouvelle méthode les exploitant.

Présentée dans la première section de ce chapitre, notre méthode consiste à construire des associations *LU française - frame* puis à filtrer les paires obtenues à l'aide du dictionnaire bilingue en exploitant la structure de FrameNet. Nous intéressant aux ressources françaises, nous n'évaluons ici que la traduction vers le français.

La ressource que nous générons n'étant pas exhaustive, nous verrons également dans la deuxième section une méthode que nous proposons pour enrichir le FrameNet.Français obtenu avec de nouvelles unités lexicales.

## 5.1 Traduction de FrameNet

Cette section est consacrée à la traduction des unités lexicales de FrameNet vers le français. Nous précisons dans un premier temps la méthode d'extraction des associations *LU française - frame*. Nous définissons ensuite divers scores de confiance que l'on attribue aux associations afin de ne conserver que les meilleures. Enfin, nous présentons le protocole d'évaluation et les résultats de l'extraction et du filtrage.

### 5.1.1 Extraction des associations *LU française - frame*

À l'instar de [Padó & Lapata 2005] et [Padó & Pitel 2007], nous supposons que l'anglais et le français sont deux langues suffisamment proches pour considérer que les frames et rôles sont d'un niveau sémantique suffisamment élevé pour être valide sur ces deux langues. Ainsi, nous conservons la structure originale de FrameNet pour la transposition de la ressource au français. Nous conserverons également les noms anglais donnés au frames, considérant ainsi ces étiquettes comme des symboles logiques que nous ne ferons pas intervenir en tant que langue naturelle dans nos travaux.

Nous utilisons deux dictionnaires bilingues pour traduire les unités lexicales de FrameNet, à savoir : le Wiktionnaire [4] (ressource communautaire) et *SCI-FRAN-EuRADic* un dictionnaire constitué et évalué par des linguistes dans le cadre du projet EuRADic et distribué par la société ELDA [5]. Nous parlerons désormais du dictionnaire EuRADic pour référer à celui-ci.

Le Wiktionnaire est un ensemble de dictionnaires collaboratifs multilingues. Il contient autant de parties qu'il existe de langues d'édition. Le principal avantage des ces dictionnaires par rapport aux dictionnaires traditionnels est inhérent à leur nature. Ils sont enrichis chaque jour par leurs utilisateurs et les néologismes y sont ainsi d'autant plus rapidement ajoutés. Nous travaillons avec une version datée du 20 janvier 2009 pour le Wiktionnaire français et du 3 février 2009 pour le Wiktionnaire anglais. À ces dates, la ressource française contenait 1 194 408 pages et la ressource anglaise 1 209 371. Parmi ces entrées, seul un nombre restreint propose des traductions. Celles-ci nous permettent d'extraire un total de 27 109 paires de traductions.

EuRADic et le Wiktionnaire prennent tous les deux en compte les *expressions multimots*†. Ceci était aussi le cas des ressources utilisées par [Padó & Pitel 2007], nous ne perdons donc pas cette partie du vocabulaire. De plus les Wiktionnaires partagent une caractéristique que nous souhaitons mettre en avant : les traductions sont souvent catégorisées en fonction des différents sens distingués pour les mots. Nous verrons dans la section suivante comment cette particularité peut aider à filtrer les données.

Nous construisons indépendamment deux ressources distinctes, l'une construite avec les paires de traduction trouvées dans le Wiktionnaire, l'autre avec les paires de traduction d'EuRADic. L'idée est d'associer aux frames les unités lexicales françaises qui sont des traductions des unités lexicales anglaises que ces frames possèdent.

Après une extraction brute effectuée à l'aide des dictionnaires bilingues, nous obtenons d'une part 19 912 paires *LU française - frame* provenant des 27 109 paires de traduction présentes dans

4. `http://wiktionary.org`
5. ELDA - Evaluations and Language resources Distribution Agency: http://www.elda.org/

le Wiktionnaire et d'autre part 57 787 paires *LU française - frame* issues des 243 539 paires de traduction présentes dans EuRADic.

En effet comme le montre la figure 5.2 où les traductions réalisées de l'anglais vers le français sont matérialisées par les flèches pleines, seules les unités lexicales présentes dans FrameNet sont traduites. Ceci explique pourquoi le nombre de paires *LU - frame* obtenu est inférieur au nombre de paires de traduction. Comme on le voit également sur cette figure, les unités lexicales anglaises peuvent posséder plusieurs traductions, ce qui explique pourquoi le nombre de paires *LU - frame* obtenus après traduction est plus important qu'au départ.

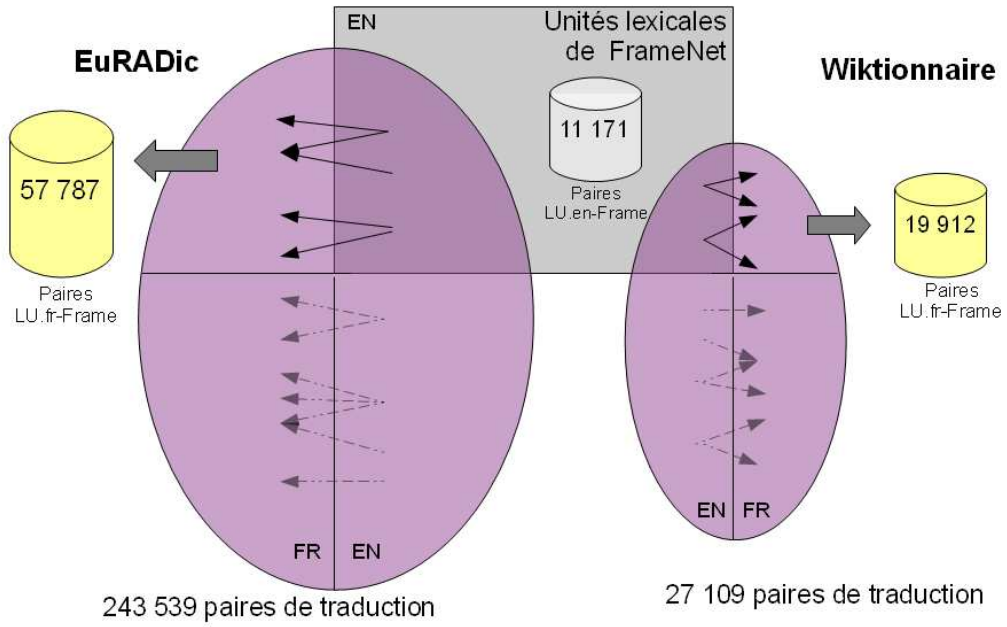



FIGURE 5.2 – Extraction des paires *LU française-frame*

### 5.1.2 Filtrage : définition des scores

Les unités lexicales obtenues dans la langue cible ne sont pas parfaites, principalement en raison de la polysémie des mots source anglais. Par exemple, l'unité lexicale anglais *depression* présente dans les frames *Medical_conditions* et *Natural_features* est traduite par *dépression*, *abattement* et *mélancolie*. L'unité *dépression* est attribuée correctement aux deux frames *Medical_conditions* et *Natural_features*. En revanche, les unités lexicales *abattement* et *mélancolie* sont attribuées toutes les deux à ces mêmes deux frames, ce qui est correct pour *Medical_condition* mais devient une erreur pour *Natural_features*.

Dans le but de ne conserver que les associations correctes, nous proposons pour chaque paire *LU-frame*, différents scores de confiance en deçà desquels les traductions ne seront pas conservées. Nous étudions alors l'influence de ce filtrage sur la qualité de la ressource obtenue.

Nous définissons tout d'abord le score $S_1$ à partir duquel nous dérivons cinq autres scores $S_i$ fondés sur différentes heuristiques. Les deux premiers utilisent la structure du FrameNet source tandis que les trois derniers utilisent les données produites après traduction.

#### 5.1.2.1 Score initial

Soit $t$ une unité lexicale cible. Soit $f$ une frame à laquelle elle est attribuée. Soit $sourcelusInFrame(f)$ l'ensemble des unités lexicales source (anglaises) de la frame $f$. Soit $sources(t, f) \subseteq sourcelusInFrame(f)$ le sous-ensemble de $sourcelusInFrame(f)$ ayant pour traduction $t$. $S_1(t, f)$ est alors la somme des nombre de sens de chaque élément de $sources(t, f)$. Le nombre de sens est toujours de 1 pour les traductions issues d'EuRADic et il peut varier pour les traductions issues du Wiktionnaire puisque les traductions sont classées en fonction de leur sens. Nous parlerons de score de traduction ($scoreTrad$) pour faire référence à ce nombre. Ainsi $S_1$ est défini de la façon suivante :

$$S_1(t, f) = \sum lu \in sources(t, f) scoreTrad(lu, t)$$

Par exemple, on voit dans le tableau 5.1 que pour la frame *Ingestion*, les dictionnaires bilingues permettent de produire les paires *drink.v* - *boire.v* et *quaff.v*-*boire.v*. La paire *drink.v* - *boire.v* obtient un score de traduction de 2, parce que *boire* est une traduction de *drink* à la fois pour le sens *consume liquid through the mouth* et le sens *consume alcoholic beverages*. La paire *quaff.v*-*boire.v* obtient un score de traduction de 1 car *quaff* n'a qu'un seul sens dans le Wiktionnaire. Ainsi, le score $S_1$ de *boire.v* - *Ingestion* est la somme de 2 pour la paire *boire.v* - *drink.v* et de 1 pour la paire *boire.v* - *quaff.v*, soit un total de 3.

| |
|---|
| remettre.v {put back.v: 1} 1 |
| **boire.v {quaff.v: 1, drink.v: 2} 3** |
| alimenter.v {feed.v: 1} 1 |
| déjeuner.v {lunch.v: 1, dine.v: 1, feed.v: 1, eat.v: 1} 4 |
| (...) |

TABLE 5.1 – Échantillon de scores pour la frame *Ingestion*

De plus, en raison de la structure des Wiktionnaires, il y a quatre possibilités distinctes pour extraire des paires de traductions comme cela est montré dans le tableau 5.2. Quand une paire de traduction apparaît dans différents endroits du Wiktionnaire, nous utilisons le score maximal produit par cette paire dans ces localisations. Nous considérons que cela correspond mieux aux nombres de

sens possibles de l'unité lexicale que la somme des scores car ces différentes localisations apportent nécessairement une redondance d'information.

Revenons sur notre exemple de *boire.v* dans la frame *Ingestion*. *boire.v* est une traduction de *drink.v* avec un score de traduction de 2 dans le Wiktionnaire anglais à l'entrée *drink*, puisque cette traduction est présente pour les deux sens distincts *consume liquid through the mouth* et *consume alcoholic beverages*. Le score de traduction de cette paire est de 2 même si elle a un score de traduction de 1 dans les trois autres localisations, les sens n'ayant pas été distingués dans celles-ci.

| Langue Wiktionnaire | Entrée du Wiktionnaire | Langue source | Langue cible |
|---|---|---|---|
| Français | boire | Français | Anglais |
| Français | drink | Anglais | Français |
| Anglais | boire | Français | Anglais |
| Anglais | drink | Anglais | Français |

TABLE 5.2 – Quatre possibilités d'extraction de traductions dans le Wiktionnaire

Si une unité lexicale est une traduction de plusieurs unités lexicales source de la frame donnée, alors elle a plus de chance d'être une bonne traduction pour cette frame. Ainsi, nous conservons les paires *unité lexicale - frame* dont le score dépasse un certain seuil. Ce seuil sera par apprentissage par l'utilisation de données d'entraînement, comme nous le verrons à la section 5.1.3.3.

#### 5.1.2.2 Scores structurels

**Score S2** Si une unité lexicale source est polysémique (c'est-à-dire si elle est présente dans plus d'une frame), il y a plus de risques que ses traductions soient erronées. En conséquence, nous introduisons le score $S_2$ diminuant le score $S_1$ en fonction du nombre de frames contenant les unités lexicales source. Pour ne pas être filtrées, les unités lexicales cible issues d'unités lexicales source polysémiques doivent donc avoir un score initial $S_1$ assez élevé, ce qui n'est pas requis pour les unités lexicales cible issues d'unités lexicales source monosémiques.

$$S_2(t,f) = \frac{S_1(t,f)}{|\{F \in FrameNet / \exists s \in sources(t,f), s \in F\}|^{\alpha}}$$

Ainsi le score $S_1$ de chaque LU (unité lexicale) traduite est divisé par le nombre de frames contenant la LU source. Ce nombre est élevé à une puissance $\alpha$, ce qui permet de moduler l'impact de ce filtre sur le score.

Prenons la LU anglaise *rise.v* pour exemple. Elle est présente dans 31 frames différentes et est donc très polysémique. Elle est présente notamment dans la frame *Getting_up*. Deux de ses traductions sont *augmenter.v* et *se lever.v*. Lorsque nous augmentons le seuil d'acceptation, les traductions

ont besoin d'avoir un score $S_1$ plus élevé pour pouvoir être conservées après filtrage. C'est le cas pour la LU *se lever.v*, qui est aussi une traduction de l'anglais *get up.v* apparaissant dans la frame *Getting_up* correspondant à l'action de *quitter son lit*. N'ayant pas d'autre traduction anglaise dans cette frame, *augmenter.v* sera alors éliminé comme on le souhaitait.

**Score S3** Le score $S_1$ nous indique la confiance que l'on peut avoir en une traduction dans une frame donnée. Plus la frame a de LUs source, plus il est probable que différentes LUs source donnent une même traduction, et par conséquent plus il est probable que les scores de traduction $S_1$ soient incrémentés artificiellement. Nous pouvons alors être moins tolérant pour ces frames qui ont beaucoup de LUs source. Ceci est réalisé par le score $S_3$.

$$S_3(t, f) = \frac{S_1(x, f)}{|sourcelusInFrame(f)|^{\alpha}}$$

Comme dans le cas du score $S_2$, un coefficient $\alpha$ nous permet de modifier l'influence du score.

Illustrons à nouveau cette heuristique par un exemple. La frame *Container* a 119 *LUs* anglaises. Il est donc très probable qu'elle contienne des synonymes, ce qui augmente également la probabilité que plusieurs LUs source produisent une/des LU(s) cible commune(s) et ainsi qu'un score $S_1$ élevé soit attribué à une LU française dans cette frame. C'est ce qui se produit avec la LU française *bac.n* qui est effectivement la traduction de 15 des 119 *LUs* anglaises (jar.n, bucket.n, chest.n, bottle.n, case.n, can.n, pail.n, urn.n, container.n, crate.n, sack.n, pot.n, tin.n, box.n, jug.n). Avec le score $S_3$, le système devient plus exigeant avec les LUs françaises contenues dans des frames dont le nombre de LUs source est élevé.

#### 5.1.2.3 Scores cible

**Score S4** En approfondissant la dernière idée développée dans la section précédente, nous pouvons dire que chaque traduction d'une LU source favorise la probabilité d'incrémentation du score $S_1$, et plus particulièrement quand une LU source produit plusieurs traductions.

Nous pouvons donc considérer non plus le nombre de LUs source, mais le nombre total de traductions produites dans la frame donnée (redondance comprise). C'est ainsi que nous définissons le score $S_4$ de la façon suivante :

$$S_4(t, f) = \frac{S_1(t, f)}{|bilingualTrans(f)|^{\alpha}}$$

avec bilingualTrans(f) l'ensemble des paires de traduction extraites des dictionnaires bilingues pour la frame $f$.

**Score S5** Nous pouvons aussi considérer que moins une frame a de LUs cible, plus celles-ci sont importantes et nous ne souhaitons pas les filtrer à outrance. L'objectif du filtre $S_5$ est de réduire

l'exigence demandée aux LUs présentes dans les frames contenant peu de LUs cible afin d'augmenter leur rappel.

Ainsi, le score $S_1$ de chaque LU cible est divisé par le nombre de LUs cible dans la frame donnée pour produire le score $S_5$ :

$$S_5(t, f) = \frac{S_1}{targetlusInFrame(f)^{\alpha}}$$

avec $targetlusInFrame(f)$ l'ensemble des unités lexicales cibles de la frame $f$.

Les conséquences de ce filtrage pourraient être discutées. En effet, dans les frames de grande taille, le système ne conserverait que les LUs de meilleurs scores et abandonnerait les plus mauvaises (avec les faux négatifs que cela implique) mais en conservant une bonne précision. En revanche, dans les frames de petite taille, le système aurait tendance à être plus indulgent aux dépends de la précision mais avec l'espoir que l'augmentation du rappel attendue soit plus importante que la perte de précision. Nous verrons dans l'évaluation de nos filtres que cette heuristique apporte néanmoins un gain de performance.

**Score S6** Plus une LU cible est présente dans beaucoup de frames, moins elle véhicule de sens dans une frame donnée. C'est le cas de certains verbes de fréquence très élevée comme *être*, *avoir*, *mettre*, etc. Le score $S_6$ tente de prendre en compte cette notion en diminuant les scores $S_1$ des LUs cibles apparaissant dans beaucoup de frames.

Le score $S_1$ de chaque LU cible est divisé par le nombre de ses occurrences dans toutes les frames du FrameNet cible pour donner le score $S_6$ :

$$S_6(t, f) = \frac{S_1}{|\bigcup_{f_i \in FrameNet} cibles(t, f_i)|^{\alpha}}$$

Nous prenons ici pour exemple la LU *rue.n* dont le score S1 est de 1 dans la frame *Roadways* en tant que traduction de *street.n* et dont le score S1 est également de 1 pour la frame *Mesure_linear_extent* en tant que traduction de *block.n*. On aura ainsi : $S6(rue.n, Roadways) = \frac{1}{2} = 0.5$.

### 5.1.3 Évaluation des filtres et de la ressource traduite

Comme ces scores emploient des paramètres qui restent à fixer, nous avons produit un ensemble de développement servant à optimiser les paramètres ainsi qu'un ensemble de test pour évaluer les productions finales.

#### 5.1.3.1 Critères d'évaluation

Les indicateurs qui nous intéressent sont très classiques, il s'agit de la *Précision*, du *Rappel*, de la $F - mesure$ et de la F-mesure pondérée $F_{0,5} - mesure$. Les poids de précision et rappel sont

égaux pour la $F - mesure$ tandis que la précision est favorisée dans le cas de la $F_{0,5} - mesure$. Les mesures sont définies respectivement ci-dessous :

$$P = \frac{Nombre\_de\_LUs\_correctes}{Nombre\_de\_LUs\_présentes}$$

$$R = \frac{Nombre\_de\_LUs\_correctes}{Nombre\_de\_LUs\_correctes\_dans\_la\_Vérité\_terrain}$$

$$F_\beta = (1+\beta^2)\frac{PR}{\beta^2 P + R} \qquad (F = F_1)$$

Les mesures sont calculées pour chaque frame et c'est la moyenne qui est prise en compte. Cela permet de donner une importance équivalente à toutes les frames indépendamment de leur taille.

#### 5.1.3.2 Données de développement et données de test

Les deux ensembles de développement et de test sont produits en réunissant une ressource dite globale par l'union des trois ressources suivantes : FrameNet.Fr ([Padó & Pitel 2007]), la ressource non filtrée construite à l'aide des traductions du Wiktionnaire et la ressource non filtrée construite à l'aide des traductions d'EuRADic.

Pour l'ensemble de développement, nous avons choisi un échantillon de 10 frames de telle sorte que leur nombre d'unités lexicales soit représentatif de la distribution globale (quantiles). Nous avons corrigé manuellement les frames correspondantes dans la ressource globale en enlevant les unités lexicales jugées incorrectes. L'ensemble résultant est alors considéré comme vérité-terrain avec laquelle les frames de chaque ressource filtrées pourront être comparées.

En ce qui concerne l'ensemble de test, nous avons sélectionné les 10 frames utilisées par [Padó & Pitel 2007]. L'ensemble de test a été obtenu en appliquant le procédé employé précédemment pour l'ensemble de développement.

Nous mentionnons ici le fait que les mesures utilisées évaluent plus la qualité des filtres que de la ressource elle-même. Le rappel est notamment biaisé car l'évaluation des ressources filtrées se fonde sur la ressource globale annotée comme vérité-terrain. On considère donc qu'il ne peut y avoir d'autres unités lexicales que celles produites par la transposition des unités lexicales anglaises (ou présentes dans FrameNet.Fr). Les mesures utilisées permettent néanmoins d'optimiser nos paramètres pour obtenir le meilleur filtrage possible en maximisant $F_{0,5} - mesure$.

#### 5.1.3.3 Paramétrage des filtres

En jouant sur les paramètres des filtres, nous pouvons produire des ressources aux propriétés différentes. Nous essayons de construire une ressource d'une taille raisonnable tout en conservant une bonne précision. Afin d'obtenir un tel résultat, nous faisons varier le paramètre $\alpha$ de chaque score

et conservons celui qui maximise la $F_{0,5} - mesure$ sur l'ensemble de développement, favorisant ainsi plus particulièrement la précision. Nous préférons privilégier la précision plus que le rappel puisque le nombre d'unités lexicales dans la ressource non filtrée est très élevé par rapport à la taille du FrameNet original. Les paramètres que nous faisons varier sont d'une part les différents $\alpha$ employés dans chaque score $S_i$ et d'autre part le seuil en-dessous duquel les unités lexicales sont éliminées car considérées non fiables.

Pour chaque dictionnaire et chaque score utilisé, nous rapportons dans le tableau 5.3 les résultats pour le paramètre $\alpha$ ayant donné la meilleure $F_{0.5} - mesure$. Nous fixons ainsi les paramètres $\alpha$ des différents scores.

| **Ressource** | $\alpha$ | $P$ | $R$ | $F_{0,5}$ |
|---|---|---|---|---|
| Wiktio | | 63% | 40% | 53% |
| $+S_1$ | | 63% | 40% | 53% |
| $+S_2$ | 1 | 65% | **40%** | 54% |
| $+S_3$ | 1 | 63% | 40% | 53% |
| $+S_4$ | 0.5 | 66% | 38% | 53% |
| $+S_5$ | 0.75 | 66% | 38% | 53% |
| $+S_6$ | 1 | 70% | 36% | **55%** |
| EuRADic | | 51% | 93% | 56% |
| $+S_1$ | | 74% | 34% | 58% |
| $+S_2$ | 0.75 | 59% | **75%** | 60% |
| $+S_3$ | 0.25 | 69% | 51% | 59% |
| $+S_4$ | 0.1 | 71% | 46% | 60% |
| $+S_5$ | 0.25 | 71% | 46% | 60% |
| $+S_6$ | 0.25 | 68% | 55% | **64%** |

TABLE 5.3 – Paramétrage des filtres par maximisation de la $F_{0,5} - mesure$

Ce tableau montre que l'utilisation des filtres présente un comportement très différemment en fonction du dictionnaire utilisé. Ceci peut être expliqué par le fait qu'EuRADic ne comporte pas de distinction de sens. Toutes les paires de traduction génèrent donc des scores de traduction de 1, ce qui n'est pas le cas pour les paires de traduction issues du Wiktionnaire.

Nous pouvons cependant remarquer un certain nombre de points communs aux deux dictionnaires : $S_2$ permet de produire la ressource avec le meilleur rappel tandis que $S_6$ permet de produire la ressource avec la meilleure $F_{0,5} - mesure$.

#### 5.1.3.4 Combinaison des filtres

Maintenant que nous avons optimisé nos paramètres $\alpha$, nous pouvons les fixer afin de combiner ces différents filtres. Pour maximiser l'effet des filtres, nous avons choisi de les combiner linéairement. Pour cela, les scores sont tous normalisés avec un écart-type de 1 et une moyenne de 0.

| | Combinaison linéaire | Toutes les frames | | Ensemble de test | | | |
|---|---|---|---|---|---|---|---|
| Ressource | | #LU-frames | #Frames | $P$ | $R$ | $F_{0,5}$ | $F$ |
| **FrameNet** | | **11 171** | **796** | | | | |
| Wi_sansfiltre | | 19 912 | 781 | 70% | 33% | 57% | 44% |
| Wi_P095 | $\frac{1}{4}.S_2 + \frac{1}{4}.S_5 + \frac{1}{2}.S_6$ | 2 889 | 686 | 94% | 11% | 33% | 18% |
| Wi_F05max | $\frac{1}{4}.S_1 + \frac{1}{2}.S_4 + \frac{1}{4}.S_6$ | **15 720** | **781** | **74%** | **30%** | **56%** | **42%** |
| EuRADic_sansfiltre | | 57 787 | 795 | 58% | 84% | 61% | 67% |
| EuRADic_P095 | $\frac{3}{4}.S_2 + \frac{1}{4}.S_6$ | 616 | 210 | 100% | 2% | 10% | 4% |
| EuRADic_F05max | $\frac{1}{4}.S_2 + \frac{3}{4}.S_6$ | **24 885** | **767** | **74%** | **44%** | **63%** | **53%** |
| **FrameNet.Fr** | | **6 659** | **480** | **77%** | **23%** | **43%** | **31%** |
| *Union* | | | | | | | |
| $Wi \cup EuRADic$ | | 65 488 | 796 | 57% | 92% | 61% | 69% |
| $W\_P_{0.95} \cup E\_P_{0.95}$ | | 3 256 | 695 | 94% | 12% | 35% | 20% |
| $W\_F_{0.5max} \cup E\_F_{0.5max}$ | | **34 121** | **793** | **70%** | **59%** | **67%** | **63%** |
| *Intersection* | | | | | | | |
| $Wi \cap EuRADic$ | | **12 211** | **773** | **82%** | **25%** | **56%** | **38%** |
| $Wi\_F_{0.5max} \cap E\_F_{0.5max}$ | | **6 484** | **724** | **95%** | **15%** | **43%** | **25%** |
| Combinaison avec FrameNet.fr | | 7 814 | 742 | 95% | 18% | 49% | 29 % |

TABLE 5.4 – Évaluation de différentes sources sur l'ensemble de test

$$Score = \sum_{i=\in[|1..6|]} w_i.S_i$$

$$avec\ w_i \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}\ et\ \sum_i w_i = 1$$

Nous avons procédé à une variation systématique des poids de la combinaison linéaire et du seuil de filtrage pour produire la ressource avec la meilleure $F_{0,5} - mesure$ possible sur l'ensemble de développement et ainsi fixer les poids optimaux $w_i$ ainsi que le seuil de filtrage.

Nous construisons également une ressource robuste avec une précision arbitraire de 95% à 0,05 près ($P_{0,95}$). Nous optimisons alors les paramètres de telle sorte que la ressource produite obtienne le meilleur rappel possible pour une précision comprise dans l'intervalle $[0,945, 0,955]$.

#### 5.1.3.5 Analyse des résultats

Les résultats rapportés dans le tableau 5.4 sont calculés sur les 10 frames de l'ensemble de test utilisé par [Padó & Pitel 2007] dans leur propre étude.

Pour les deux types de ressources que nous avons souhaité construire, chacun des scores définis

a joué un rôle significatif dans le filtrage, excepté le score $S_3$. Nous avons aussi fixé les paramètres en optimisant la $F_{0,25} - mesure$. Il s'est avéré que les meilleurs paramètres $\alpha$ et poids de la combinaison linéaire sont légèrement différents. Nous avons également remarqué qu'alors le score $S_3$ était utilisé dans certaines des meilleures combinaisons. Ces résultats montrent que nous devons entraîner les paramètres des scores et de la combinaison linéaire chaque fois que nous voulons produire une ressource avec des caractéristiques différentes, en fonction de l'application qui utilisera cette ressource.

En comparaison aux scores utilisés indépendamment, le filtrage s'avère plus efficace quand les scores sont combinés et les résultats montrent que les six scores définis sont utiles pour améliorer la précision des ressources traduites, validant ainsi nos heuristiques.

En ce qui concerne les ressources produites, nous trouvons plusieurs résultats intéressants : les ressources obtenues en maximisant la $F_{0,5} - mesure$ sont d'une précision acceptable (74% à la fois pour la ressource construite à partie du Wiktionnaire et pour celle construite à partir d'EuRADic, contre 77% pour FrameNet.Fr) tandis qu'elles constituent déjà un nombre plus important de paires que le FrameNet de Berkeley (respectivement 15 720 et 24 885 contre 11 171).

Nous montrons également que nous pouvons obtenir une précision estimée à 95% pour un FrameNet couvrant 724 des 796 frames du FrameNet de Berkeley, en effectuant l'intersection des ressources construites à partir d'EuRADic et du Wiktionnaire avec $F_{0,5}$ maximisée (la tentative précédente de FrameNet.Fr ne concernait que 480 des frames pour une précision estimée plus faible).

La dernière ligne du tableau est ajoutée pour information et présente un résultat produit par l'union des intersections de ressources par paires ($Wi\_F_{0,5}max \cap EuRADic\_F_{0,5}max$, $Wi\_F_{0,5}max \cap FrameNet.Fr\_sansfiltre$ et $FrameNet.Fr\_sansfiltre \cap EuRADic\_F_{0,5}max$), ce qui permet de conserver une ressource dont la précision est estimée à 95% tout en augmentant de 20 % la taille de la ressource.

Finalement, nous remarquons que le gain en précision dû au filtrage est plus important pour la ressource construite avec EuRADic (58% à 74%), mais cela est fait au détriment de la taille de la ressource : $Eu\_F_{0,5}$ est réduite à moins de la moitié de sa taille originale tandis que $Wi\_F_{0,5}$ conserve environ 80% de sa taille originale. Comme le nombre de paires de traduction présentes dans EuRADic est beaucoup plus élevé que pour le Wiktionnaire, EuRADic permet néanmoins de produire la ressource la plus grande. Cependant, la différence dans la taille des ratios de réduction nous invite à penser que les filtres sont plus adaptés à la structure du Wiktionnaire et qu'un Wiktionnaire contenant autant d'entrées qu'EuRADic permettrait certainement d'obtenir une ressource encore meilleure que celle obtenue avec EuRADic.

Nous voyons deux raisons expliquant pourquoi la ressource obtenue avec le Wiktionnaire est filtrée plus efficacement. La première est due à la structure plus complexe du Wiktionnaire qui catégorise les traductions en fonction du sens des mots. Cela permet de mieux prendre en compte

la polysémie lors du filtrage. La deuxième raison provient de la non-exhaustivité des ressources. En effet, en général les unités lexicales sont présentes dans FrameNet dans leur sens le plus courant (peut-être car la ressource est toujours en cours de construction). Dans le Wiktionnaire, c'est assez similaire, les sens et usages les plus courants sont plus souvent traduits que les usages plus rares. Si une unité lexicale source de FrameNet est traduite par une unité lexicale cible correspondant à un sens rare présent dans le dictionnaire mais pour lequel l'unité lexicale source ne figure pas dans FrameNet, cela génère une erreur qui ne sera pas nécessairement filtrée. Le fait que le Wiktionnaire ne contienne que des sens courants permet d'éviter ce genre de biais.

### 5.1.4 Premières conclusions

Nous montrons qu'avec la méthode proposée, nous parvenons à obtenir une ressource beaucoup plus grande que celle qui existait précédemment et de meilleure qualité du point de vue de la traduction des unités lexicales.

Notre approche peut utiliser n'importe quel dictionnaire bilingue, que celui-ci fasse la distinction entre les sens ou non. Elle est théoriquement transposable à toute autre langue mais nous devons admettre que la partie de notre étude utilisant le Wiktionnaire exploite la partie française de celui-ci [6], celle-ci étant la plus développée [7]. En effet, en 2005 le Wiktionnaire français a été enrichi automatiquement grâce à des dictionnaires libres et la communauté est restée très active pour l'enrichir manuellement.

Notre méthode de traduction permet de traiter les unités lexicales françaises (de langue cible) uniquement si leurs équivalents anglais sont présents dans la ressource originale. Dans le cas contraire, c'est-à-dire si leurs équivalents anglais ne sont pas dans FrameNet, nous n'avons aucun moyen d'attribuer une frame aux termes français. Ainsi notre traduction ne peut pas produire une ressource exhaustive d'une part du fait que la ressource originale de Berkeley est construite manuellement et par ailleurs toujours en construction, d'autre part en raison du filtrage appliqué lors de la traduction des mots présents. C'est pourquoi nous étudions également l'enrichissement potentiel de notre ressource traduite.

6. urlhttp://wiktionary.org

7. En mars 2010, la partie française contient plus de 1 659 000 entrées, la partie anglaise plus de 1 609 000.
Seules 18 langues contiennent plus de 100 000 entrées (par ordre décroissant): Français, Anglais, Lituanien, Turc, Chinois, Russe, Vietnamien, Ido (langue issue de l'Esperanto), Polonais, Portugais, Finnois, Hongrois, Norvégien, Tamoul (langue du Sud de l'Inde), Allemand, Italien, Suédois, Grec.
Nous pouvons aussi remarquer que la partie allemande contient seulement 104 294 entrées mais représente 12,8% du trafic contre 9,7% pour la partie française !

## 5.2 Enrichissement de la ressource lexicale française

Le fait que le nombre de nos paires (*unité lexicale, frame*) puisse être trois fois plus élevé que le nombre de ces paires présentes dans le FrameNet original de Berkeley ($W\_F_{0.5max} \cup E\_F_{0.5max}$) montre que la ressource anglaise originale est loin d'être exhaustive. Certaines unités lexicales devraient y figurer mais n'apparaissent dans aucune frame. C'est le cas par exemple de l'unité lexicale *taxonomy.n* qui devrait figurer dans la frame *Categorization*. Par ailleurs, certaines unités lexicales sont présentes mais n'appartiennent pas à toutes les frames qu'elles devraient déclencher. C'est le cas de l'unité lexicale *boom.n* qui apparaît seulement dans la frame *Sounds* alors qu'elle devrait être présente également dans la frame *Progress* pour son usage possible en tant que croissance, comme dans les expressions *baby boom* ou *business boom*.

En raison de ces absences, les traductions de ces unités lexicales manquantes sont rarement présentes dans les frames pertinentes dans la ressource cible. Nous nous penchons donc sur la tâche d'enrichissement de la ressource française.

Cette étape permet également de renforcer certaines unités lexicales obtenues lors de la phase de traduction. On trouve dans la littérature une approche assez similaire : [Pennacchiotti *et al.* 2008] proposent deux mesures de similarité pour réaliser cette tâche, la première est une mesure de similarité distributionnelle obtenue à partir de trois types d'espaces sémantiques (i.e. cooccurrences de fenêtre, cooccurrences syntaxiques et cooccurrences syntaxiques non typées). La seconde mesure exploite les liens sémantiques de WordNet.

### 5.2.1 Enrichissement à l'aide d'un classifieur k-NN multi-représenté

Dans le but d'enrichir notre ressource française, nous appliquons un classifieur à tous les noms du vocabulaire français figurant dans nos espaces sémantiques (∼45 000). Les classes auxquelles ces mots doivent être attribués sont les différentes frames, déjà constituées des unités lexicales obtenues par traduction.

Si le mot est déjà dans la ressource, il est temporairement exclu des frames auxquelles il appartient et réaffecté soit pour confirmer une attribution obtenue par la traduction, soit pour générer une nouvelle association *unité lexicale-frame* qui n'était pas présente à l'origine dans FrameNet (ou qui n'a pas été traduite).

Nous disposons des espaces sémantiques multi-représentés construits par [Grefenstette 2007] et sur lesquels nous avons appliqué une pondération par l'information mutuelle spécifique et une réduction de dimensions. Pour de plus amples détails, nous rappelons au lecteur que la constitution de ces espaces et les traitements appliqués dessus sont décrits à la section 3.1.1.2.

Nous utilisons pour la classification une réimplémentation de l'algorithme décrit par [Kriegel *et al.* 2005] déjà utilisé dans nos travaux sur la désambiguïsation (Section 4.1.2.2). Cet algorithme

est une variation du classifieur traditionnel k-NN. Le nouvel algorithme combine les sphères de voisinage k-NN de toutes les représentations en tenant compte de leur qualité. L'idée principale est qu'une sphère de faible entropie, c'est-à-dire contenant peu de classes mais beaucoup d'éléments dans chacune d'entre elles et plus particulièrement dans la classe majoritaire, est plus fiable qu'une sphère de voisinage k-NN de forte entropie, c'est-à-dire contenant des éléments de beaucoup de classes différentes et peu d'éléments dans chaque classe.

Soit $\{R_i\}$ un ensemble de représentations et $\{C_j\}$ un ensemble de classes, la règle de classification multi-représentée est définie de la façon suivante :

$$Cl_{mr}(o) = max_{j=1,\ldots,|C|}(\sum_{i=1}^{m} w_i.cv_{i,j}(o))$$

$cv_{i,j}$ représente la confiance que l'on peut avoir en la classe $C_j$ dans la représentation $R_i$. Nous modifierons une partie de la formule du vecteur de confiance lors de l'optimisation de nos paramètres et présenterons donc plus en détails le contenu de ce vecteur à ce moment-là.

$w_i$ est un terme correspondant à l'entropie de la représentation $R_i$ concernant les différentes classes. Ce terme dépend de $cv(o)$. Nous encourageons le lecteur à se référer à [Kriegel *et al.* 2005] pour des détails plus théoriques.

### 5.2.2 Paramètres

Beaucoup de choix doivent être faits afin de générer la ressource enrichie. Nous décrivons chacune de ces variations ci-dessous.

#### 5.2.2.1 Données d'apprentissage

Nous entraînons notre classifieur avec trois ensembles de données distincts : la ressource maximisant le score $F_{0,5}$ (notée $F$), la plus grande ressource de précision $P_{0.95}$ (notée $P$), et enfin l'union non filtrée des trois ressources Wi, EuRA et FrameNet.Fr (notée $U$).

#### 5.2.2.2 Sphère de voisinage

Nous faisons varier le paramètre $k$ fixant la taille de la sphère de voisinage de l'algorithme k-NN selon les trois valeurs 10, 25, 50.

#### 5.2.2.3 Seuils

Lors de notre enrichissement, nous aimerions être capables d'attribuer un nouveau mot à plusieurs frames à la fois lorsque cela est pertinent. Afin d'en donner la possibilité à l'algorithme, nous introduisons deux seuils sur le score de la règle de classification. Les classifications seront validées si leur score est supérieur à ces seuils. Le premier seuil ($s_1$) est statique. Il donne également la possibilité

de ne pas attribuer de classe à un nouvel élément si le score de la règle de classification est trop faible. Le second seuil ($s_2$) est un pourcentage du meilleur score pour chaque nouvelle unité lexicale. Si seul ce seuil est utilisé, tous les éléments sont attribués à au moins une classe.

#### 5.2.2.4 Espaces sémantiques

Nous utilisons les espaces sémantiques significatifs, c'est-à-dire ceux dont la relation syntaxique relie deux mots pleins (dont au moins l'un des deux est un nom). On exclut donc certains espaces tels que déterminant_substantif, préposition_substantif... Certains espaces sémantiques sont moins informatifs que d'autres en raison de l'insuffisance des données et des erreurs d'analyse (rareté intrinsèque de certaines relations comme par exemple l'*attribut de l'objet*). C'est pourquoi, nous appliquons le classifieur avec différents ensembles d'espaces sémantiques :

- chaque espace utilisé seul ;
- la combinaison de tous ;
- tous excepté l'espace *attribut de l'objet* (noté $all \backslash atb\_obj$) ;
- tous excepté les espaces *attribut de l'objet* et *attribut du sujet* (noté $all \backslash\ atb\_obj\_subj$) ;
- tous excepté les espaces *attribut de l'objet*, *attribut du sujet* et *complément de l'adjectif* ;
- la combinaison des espaces *complément d'objet, apposition* et la relation inverse de l'*apposition* ;
- la combinaison des trois espaces précédents, de l'espace *complément du nom* et de l'espace issu de la relation inverse (combinaison notée *5best*).

#### 5.2.2.5 Vecteur de confiance

Nous testons quatre façons différentes de calculer le vecteur de confiance. Le vecteur de confiance original proposé par [Kriegel *et al.* 2005] est défini de la façon suivante :

$$\forall j, 1 \leq j \leq |C| :$$

$$(A)\quad cv_{i,j}(o) = \frac{\sum_{u \in sphere_i(o,k) \wedge c(u)=c_j} \frac{1}{d_{norm}(o,u)^2)}}{\sum_{k=1}^{|c|} cv_{i,k}(o)}$$

Comme déjà mentionné à la section 4.1.2.2, nous remarquons que le vecteur de confiance utilisé dépend du nombre d'éléments qui appartiennent à la classe donnée ($|u \in sphere_i(o,k)) \wedge c(u) = c_j|$) dans la sphère k-NN mais il ne tient pas compte du nombre d'éléments dans chaque classe $j$. Pourtant, il est plus probable qu'un élément appartienne à la classe donnée si cette classe est plus grosse que les autres. Nous en venons ainsi à l'idée de prendre en considération le nombre d'éléments de chaque classe dans le calcul de ce vecteur de confiance et nous testons ainsi trois formules distinctes :

$$(B)\quad cvB_{i,j}(o) = \frac{cv_{i,j}}{|c_j|}(o)$$

| Ressource | Données d'apprentissage | k | $s_1$ | $s_2$ | Espaces sémantiques | Vecteur de confiance | Pondération |
|---|---|---|---|---|---|---|---|
| **EFN.1** *précision* | (U) | 10 | 0 | 0,95 | $all \backslash atb\_ obj\_ subj$ | D | oui |
| **EFN.2** *couverture* | (U) | 50 | 0 | 0,95 | 5best | D | non |

TABLE 5.5 – Meilleurs paramètres pour l'enrichissement

$$(C) \quad cvC_{i,j}(o) = \frac{cv_{i,j}}{log(1 + |c_j|)}(o)$$

$$(D) \quad cvD_{i,j}(o) = \frac{cv_{i,j}}{log(10 + |c_j|)}(o)$$

#### 5.2.2.6 Pondération

Les scores de traduction des unités lexicales peuvent être utilisés pour pondérer les distances dans le classifieur. Nous pouvons soit les utiliser, soit considérer toutes les données d'apprentissage comme équivalentes.

#### 5.2.2.7 Optimisation des paramètres

Comme nous procédons à un enrichissement, nous n'avons aucune vérité-terrain à laquelle nous comparer. Avec pour objectif d'optimiser nos paramètres, nous considérons la ressource globale (union des trois ressources non filtrées) comme ressource de comparaison. Nous faisons varier nos paramètres et évaluons chaque nouvelle ressource enrichie en calculant une approximation de la précision correspondant au nombre d'unités lexicales correctement réassignées divisé par le nombre total de paires enrichies dont l'unité lexicale est présente dans la ressource globale. Nous calculons également une mesure de couverture correspondant au nombre de paires réassignés (correctement ou non) sur le nombre total de paires dans la ressource globale.

Nous conservons la ressource avec la précision ainsi nommée la plus élevée (*EFN.1*) ainsi que la ressource avec la meilleure couverture (*EFN.2*) afin de les évaluer avec l'ensemble de test. Les paramètres donnant ces meilleurs résultats sont rapportés dans le tableau 5.5.

### 5.2.3 Résultats

Nous constituons une nouvelle vérité terrain par l'union de l'ensemble de test construit pour la tâche de traduction de FrameNet et de l'ensemble des deux meilleures ressources enrichies. Nous corrigeons manuellement les paires provenant des ressources enrichies et appartenant aux frames de test. Les résultats sont présentés dans le tableau 5.6.

Ces résultats nous semblent vraiment bons. La précision est estimée à 82% lorsque nous ajoutons 7 581 nouvelles unités lexicales, ce qui représente environ 75% de la taille du FrameNet original

| | Nombre total | | | Ensemble de test | |
|---|---|---|---|---|---|
| **Ressource** | **LU-frame** | **Nouvelles attributions**[3] | **Frames** | $P$ | $R$ |
| **FrameNet de Berkeley** | **11 171** | | **796** | | |
| **EFN.1** *précision* | 9 536 | 7 581 (79%)[2] | 295 | 82% | 7% |
| **EFN.2** *couverture* | 27 371 | 24 997 (91%)[2] | 359 | 61% | 10% |
| **TFN + EFN.1**[1] | 15 132 | 8 648 (57%)[2] | 727 | 86% | 18% |

[1] TFN + EFN.1 = $(Wi\_F_{0,5}max \cap Eu\_F_{0,5}max) \cup EFN.1$
[2] comparé à la ressource globale
[3] Nombre et proportion de nouvelles attributions

TABLE 5.6 – Évaluation des FrameNets enrichis

de Berkeley. En ce qui concerne la ressource construite en optimisant la couverture, nous devons garder à l'esprit que nous avons traité seulement les noms alors que la ressource traduite contient aussi des verbes, adjectifs, adverbes et prépositions. La proportion de noms est de 41% dans le FrameNet original. Notre ressource $EFN.2$ aurait donc un rappel sur les noms d'environ 25%. Une autre raison pour laquelle le rappel n'est pas très élevé en comparaison au nombre d'unités lexicales réassignées peut être avancée. La couverture a été maximisée sur l'ensemble de toutes les frames sans distinction pour les paires appartenant à des frames de tailles différentes alors que notre mesure de rappel est calculée sur les 10 frames séparément avant d'être combinée pour donner la moyenne. Ceci peut biaiser les résultats de telle sorte que les grandes frames aient un rappel élevé et les petites frames un rappel assez faible, produisant ainsi un score de couverture élevé et une moyenne des rappels intermédiaire. L'examen des données conforte cette idée. En effet les frames de grande taille semblent être d'autant plus enrichies comme le montre la figure 5.3, où nous présentons le nombre d'unités lexicales produites par l'enrichissement en fonction du nombre d'unités lexicales nominales présentes dans la frame. En effet, certaines frames se prêtent plus à la présence d'une liste de noms, comme par exemples les frames *Observable_bodyparts*, *Natural_features*, *Containers*, *Aggregate*, *Buildings*, *Food* ou *Clothing* (cf. tableau 5.7).

Ces enrichissements sont significatifs. Nous pouvons maintenant faire l'union de l'une de ces ressources avec une des ressources produites par la traduction. Par exemple, si nous combinons notre ressource enrichie *EFN.1* avec notre ressource traduite $Wi\_F_{0,5}max \cap Eu\_F_{0,5}max$, nous obtenons une ressource contenant 15 132 paires *unité lexicale-frame* avec une précision globale estimée à 86%.

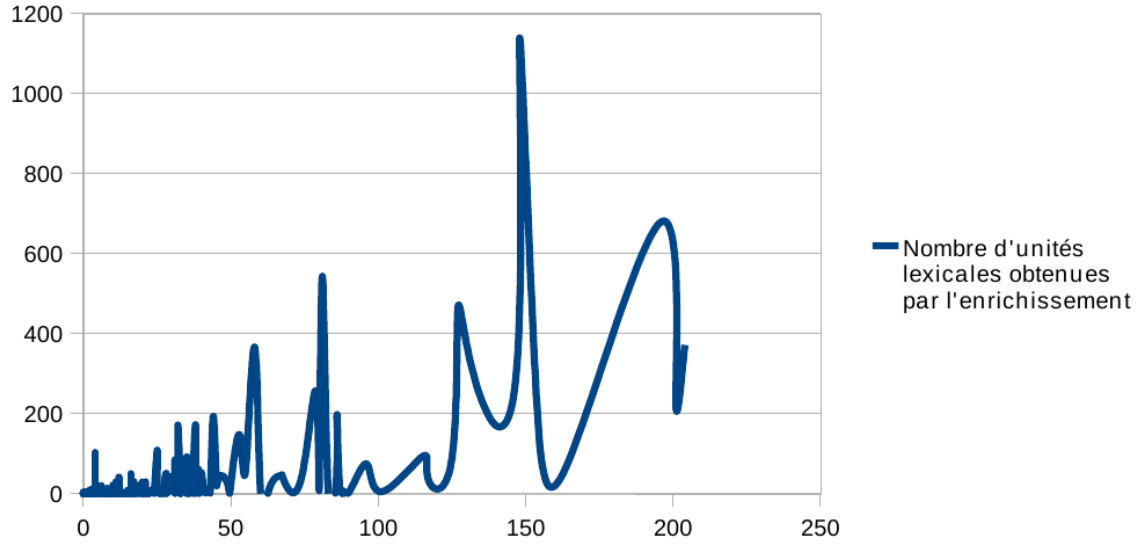



FIGURE 5.3 – Nombre d'unités lexicales produites par l'enrichissement pour chaque frame en fonction du nombre de substantifs dans les données d'apprentissage de la frame

## 5.3 Conclusions et perspectives

Nous avons proposé une nouvelle approche pour transférer le contenu de la ressource anglophone FrameNet à une autre langue et avons validé cette approche pour le français.

Nous avons obtenu deux type de ressources pour chaque dictionnaire utilisé. La première ressource est robuste (approximativement 95% de précision) mais plus petite que la ressource originale de Berkeley (58% du nombre de paires *LU-frame* de FrameNet ou 70% si nous incluons FrameNet.Fr dans notre combinaison). La seconde est plus équilibrée : elle est beaucoup plus grande que le FrameNet original de Berkeley (34 121 paires *LU-frame*, c'est-à-dire environ trois fois le nombre de paires présentes dans le FrameNet de Berkeley) avec une précision diminuée à 70% mais un rappel beaucoup plus élevé (estimé à 59%). Une comparaison avec la traduction existante de [Padó & Pitel 2007] montre que nous pouvons produire une ressource $Wi \cap EuRADic$ dont la taille est deux fois plus grosse tout en conservant une meilleure précision (estimée à 82% au lieu de 77%). En revanche, cette approche ne permet pas d'obtenir un corpus d'apprentissage complet comme c'était le cas dans les travaux de [Padó & Pitel 2007] pour lesquels toutes les annotations de rôles étaient également projetés sur les syntagmes correspondant français.

Nous avons aussi traité le problème de l'enrichissement pour pallier les problèmes de non-exhaustivité du FrameNet de Berkeley et de nos dictionnaires. Cet enrichissement pratiqué sur les noms montre des résultats très encourageants, incitant à mettre en œuvre cette stratégie également

| | |
|---|---|
| Observable_bodyparts | métatarsien, coude, narine, canine, postérieur, orteil, trompe, talon, plexus, pupille , (...) |
| Natural_features | ruisseau, plateaux, banc de sable, altiplano, envasement, crique, vallée, désert, roc, terrasse, (...) |
| Containers | coffret, pinte, fût, caisse, cruche, bouilloire, cageot, housse, bonbonne, tuyaux, (...) |
| Aggregate | escouade, quatuor, gang, patrouille, multitude, quintette, banc, régiment, horde, bande, (...) |
| Buildings | halle, habitation, mairie, rempart, stade, temple, échoppe, grange, palais, médiathèque, (...) |
| Food | chocolat, palourde, thon, munster, cornichon, fondue, rognon, aubergine, andouille, champignon, (...) |
| Clothing | cagoule, minijupe, collant, cuirasse, maillot de corps, fringue, smoking, sous-vêtement, ballerine, botte, (...) |

TABLE 5.7 – Exemples de frames contenant beaucoup de noms

pour les verbes qui sont plus souvent utilisés comme prédicats.

Les résultats de l'évaluation montrent que nos scores reflètent bien la confiance que l'on peut accorder à une assignation *LU française -frame*. Ces scores de confiance sont donc à conserver pour pouvoir être pris en compte dans des traitements utilisant notre ressource.

Enfin, nous remarquons que pour enrichir la ressource anglaise, on peut soit appliquer la même méthode d'enrichissement que celle proposée pour le français, soit, si on ne dispose pas d'espaces sémantiques anglais, le processus de traduction peut être réitéré en sens inverse pour traduire vers l'anglais les nouvelles unités lexicales françaises obtenues par l'enrichissement automatique.

Maintenant que nous disposons d'une ressource française, nous pouvons procéder à l'étude d'un système d'annotation en rôles sémantiques pour le français.

# Chapitre 6

# Annotation en rôles sémantiques

Nous proposons un travail pionnier dans l'annotation sémantique de rôles en français. Pour parvenir à nos fins, nous utilisons le FrameNet français décrit au chapitre précédent. Celui-ci possède la même structure que le FrameNet d'origine de Berkeley. Ainsi nous disposons de quasiment l'intégralité des frames définies dans FrameNet, ainsi que d'exemples d'annotations sur le corpus anglophone. Notre FrameNet français fournit les unités lexicales françaises déclenchant la présence de frames.

Dans un premier temps, nous décrivons le système étudié exploitant les informations que nous venons de mentionner. Nous procédons ensuite à l'évaluation des annotations produites par un tel système. Enfin, nous revenons sur l'intérêt de l'annotation sémantique de rôles dans un système de recherche d'information.

## 6.1 Système de SRL semi-supervisé pour le français

L'annotation en rôles sémantiques est définie comme une tâche consistant d'une part à détecter dans une phrase un ou plusieurs prédicat(s) définissant un scénario ainsi que les différents arguments réalisant les rôles de ces scénarios, et d'autre part à les identifier par rapport aux étiquettes d'une ressource de référence. Pour ce faire, nous avons choisi de travailler avec une ressource de type FrameNet. Nous avons déjà évoqué rapidement les raisons de ce choix à la section 2.2.1.3 et dans l'introduction du chapitre 5 et nous y reviendrons plus en détails dans la section suivante lorsque nous introduirons l'indexation des rôles sémantiques en recherche d'information.

Dans notre phrase exemple mentionnée en introduction et reproduite ci-dessous, une telle annotation consiste d'une part à détecter les différents syntagmes, et d'autre part à déterminer que c'est la frame *Activity_stop* déclenchée par le prédicat *s'achever* qui est réalisée par les rôles *Agent, Time* et *Result*, eux-mêmes remplis par les différent syntagmes (de même couleur sur la figure). Dans cet exemple tous les syntagmes correspondent à un rôle mais ce n'est pas souvent le cas !

| [Agent] | **Activity_stop** | [Time] | [Result] |
|---|---|---|---|
| La table ronde «radio-fréquences, santé, environnement» | **s'est achevée** | ce lundi | avec une dizaine de pistes mais sans aucune mesure forte. |

Nous avons pu relever dans notre état de l'art deux points essentiels dans l'annotation automatique de rôles sémantiques.

D'une part, la plupart des systèmes sont des systèmes d'apprentissage supervisé. Bien que ceux-ci soient ceux qui donnent les meilleurs résultats dans les campagnes d'évaluation, cela est aussi une limite dans un cadre plus général. En effet, pour d'autres langues que l'anglais il n'existe pas (ou très peu) de corpus annoté pouvant servir de données d'apprentissage. D'autre part, le traitement des cas ne figurant pas dans les corpus d'apprentissage est souvent très limité. Nous nous intéressons ici à une méthode d'annotation semi-supervisée pour le traitement du français.

D'autre part, il existe une forte corrélation entre les relations syntaxiques et les relations sémantiques [Rosen 1984], [Tenny 1994]. Il semble donc indispensable d'utiliser l'information syntaxique disponible.

Enfin, nous constatons que l'information syntaxique ne suffit pas pour deux raisons principales. La première raison est due à l'existence de différents cadres de sous-catégorisation possibles pour chaque verbe. Pour illustrer notre propos, nous reprenons les exemples de [Fillmore 1968] et présents dans l'article de [Fernandez *et al.* 2002] pour le verbe *fermer* :

1. Le vent ferme les fenêtres d'un coup.
2. La porte s'est fermée d'un coup.
3. À 7 heures, les portes ferment.
4. La porte est fermée.
5. Il a laissé la porte fermée.
6. La porte se ferme facilement.

Nous voyons dans ces différentes phrases que la fonction grammaticale ne suffit pas à déterminer le rôle sémantique d'un argument. Par exemple, le sujet du verbe *fermer* (sans usage pronominal) est *le vent* dans la phrase 1 et *les portes* dans la phrase 3. Pourtant *le vent* tient le rôle générique d'*agent* tandis que *les portes* tient un rôle générique de *thème* ou de *patient*.

La deuxième raison provient de l'existence de compléments verbaux qui remplissent une même dépendance grammaticale sans pour autant avoir le même rôle thématique. Voici quelques exemples de ce genre de cas :

1. *Marie a acheté un cartable à sa fille.*
2. *Marie a acheté un livre à un brocanteur.*

3. *Jean mange avec des amis.*
4. *Jean mange avec des baguettes.*

Dans ces phrases, nous voyons que les arguments *à sa fille* et *à un brocanteur* des phrases 1 et 2 sont tous les deux compléments d'objet indirect de *acheter*, cependant *à sa fille* tient un *rôle* de *destinataire/bénéficiaire* tandis que *à un brocanteur* tient un rôle de *source/vendeur*. L'exemple des phrases 3 et 4 est similaire : les arguments *avec des amis* et *avec des baguettes* sont tous les deux compléments circonstanciels (employés avec la même préposition) mais *avec des amis* tient un rôle de *co-agent* tandis que *avec des baguettes* tient un rôle d'*instrument*.

Pour les deux phénomènes décrits ici, il devient indispensable d'exploiter également une information lexicale externe afin de pouvoir déterminer quels arguments remplissent quels rôles sémantiques.

Le système que nous développons fait appel aux espaces sémantiques déjà utilisés au cours de notre travail. Ces espaces sont construits à partir de cooccurrences syntaxiques. Ils vont nous permettre d'exploiter d'une part les liens entre informations lexicales et rôles sémantiques, et d'autre part les liens entre informations syntaxiques et rôles sémantiques.

### 6.1.1 Vue générale du système

La figure 6.1 décrit le fonctionnement général de notre algorithme. L'entrée de notre système est un texte annoté en dépendances syntaxiques par l'analyseur LIMA [Besançon & Chalendar (de) 2005]. La sortie est le même texte annoté cette fois-ci en dépendances sémantiques, c'est-à-dire en rôles de FrameNet.

Soit $\phi$ la phrase annotée en dépendances syntaxiques que l'on doit analyser. La première phase consiste à détecter les prédicats présents dans la phrase, ainsi que les têtes de syntagmes que nous supposons arguments de ces prédicats.

#### 6.1.1.1 Détection des prédicats

Un prédicat est détecté à chaque fois qu'un terme $t$ de la phrase $\phi$ appartient à une frame de notre FrameNet français constitué au chapitre précédent. Une ou plusieurs frame(s) sont alors candidates pour l'annotation du prédicat détecté.

Pour chaque prédicat détecté, on instancie alors un ensemble de structures candidates noté $\mathcal{C}$. Cet ensemble contient autant de structures que le prédicat déclenche de frames. Chaque structure candidate contient le nom de la frame déclenchée ainsi que l'ensemble des noms de rôles sémantiques principaux (*Core*) qu'elle peut voir réalisés.

Soit $p$ un prédicat détecté. On a alors :

$$\mathcal{C} = \{\forall f \in frames(p), f \oplus roles(f)\}$$

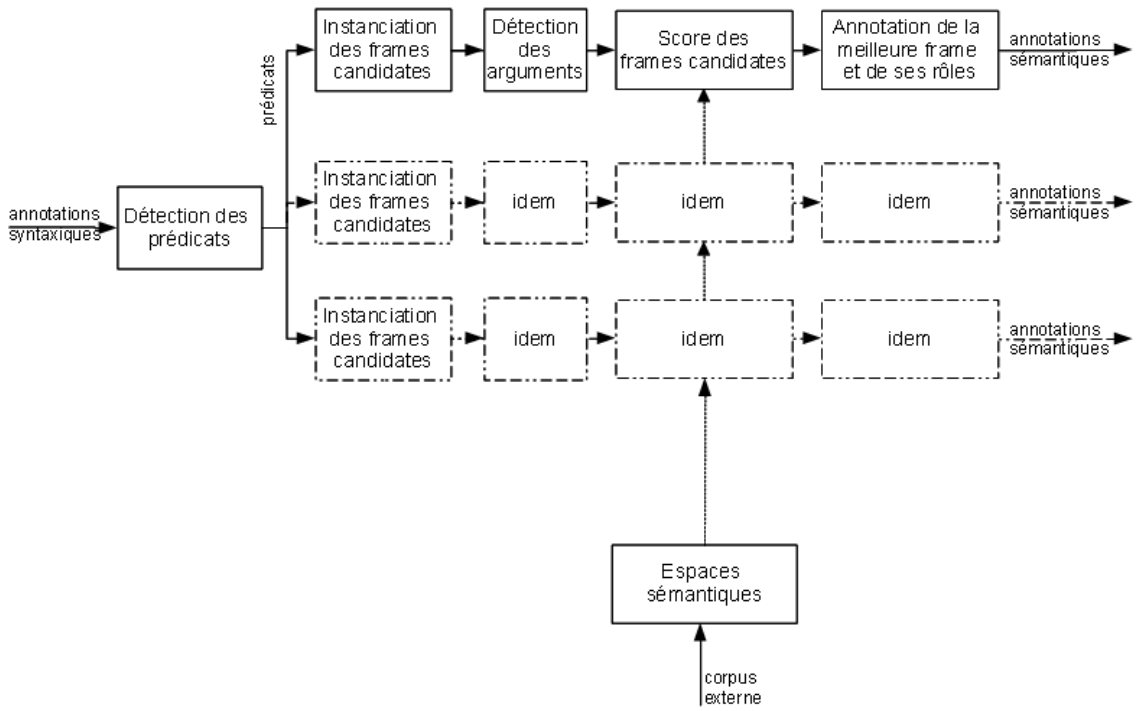



FIGURE 6.1 – Vue générale de notre algorithme d'annotation de rôles sémantiques

avec $frames(p)$ l'ensemble des frames déclenchées par le prédicat $p$, $roles(f)$ l'ensemble des rôles définis pour la frame $f$ et $\oplus$ un opérateur marquant l'association dans la structure entre $f$ et ses rôles.

Pour le prédicat *achever.v* dans notre phrase exemple, on aurait alors :

| Frame | Rôles Principaux | Rôles secondaires |
|---|---|---|
| Activity_stop | Agent | (Activity, Time, Result, Duration, Purpose (...)) |
| Activity_finish | Activity, Agent | (Time, Result, Duration, Purpose (...)) |
| Killing | Killer, Victim, Cause, Means, Instrument | (Manner, Purpose, Reason, Place (...)) |

Un score va ensuite être attribué à chacune de ces structures candidates afin de valider ou d'invalider par un seuil la présence de la frame dans la phrase analysée.

#### 6.1.1.2 Détection des arguments candidats

Pour chaque prédicat, le système procède alors à la détection de l'ensemble $\mathcal{A}$ des syntagmes candidats à être argument. De la même façon que [Swier & Stevenson 2004], nous choisissons d'utiliser exclusivement les têtes de syntagmes ainsi que leurs relations syntaxiques au reste de la phrase. L'information utilisée est alors équivalente à l'ensemble des traits issus du syntagme utilisés par [Gildea 2002].

Nous définissons la distance syntaxique entre deux termes comme étant la longueur du chemin en terme de relations syntaxiques pour aller d'un terme à un autre.

Pour les verbes, nous choisissons les arguments candidats comme étant tous les termes pleins étant à une distance de un ou de deux, à l'exception des termes grammaticaux (prépositions, conjonctions) mais y compris les pronoms. Parmi les termes à une distance de 1, on ne conserve que ceux dont la relation syntaxique les reliant au prédicat est gouvernée par le prédicat lui-même et éliminons ceux qui gouvernent la relation syntaxique donnée. Par exemple, dans la phrase illustrée par la figure 6.2 où le prédicat traité est *fermer.v*, le terme *oublié* gouvernant la relation syntaxique *complément verbal du verbe* est éliminé de la liste des arguments candidats.

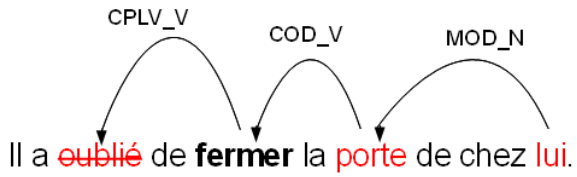



FIGURE 6.2 – Le terme gouvernant *oublié* est éliminé

On associe ensuite un poids de 2 à chaque argument à une distance de 1, ainsi qu'un poids de 1 à chaque argument à une distance de 2. Ces restrictions traduisent partiellement l'hypothèse restrictive

vérifiée par [Abend *et al.* 2009] selon laquelle la majorité des arguments des verbes appartiennent à la plus petite proposition à laquelle le verbe appartient.

Pour les noms, adjectifs et adverbes, nous ne conservons que les termes à une distance de 1 dont la relation est gouvernée par le prédicat. Enfin pour les prépositions, nous ne conservons que les termes à une distance de 1 dont la relation est gouvernée par le terme lui-même.

Pour chaque structure candidate, nous supposons pour l'instant que nous disposons d'un score traduisant la force d'association entre un argument candidat et un rôle. Ce score, calculé à partir des espaces sémantiques, sera décrit dans la section 6.1.2. Nous poursuivons pour l'instant la description générale de l'algorithme. On a donc maintenant :

$$\forall a \in \mathcal{A}, \forall r \in roles, \exists scoreAsso(a, r)$$

#### 6.1.1.3 Score des structures candidates

Pour chaque structure candidate, nous calculons toutes les bijections possibles des arguments vers les rôles, c'est-à-dire l'ensemble de tous les ensembles d'associations possibles entre les arguments candidats et les rôles. Afin de réduire la combinatoire, nous éliminons les bijections contenant des associations argument/rôle dont le score est inférieur à un certain seuil. Un score est ensuite attribué à chaque bijection restante, il correspond à la somme des scores de chaque association argument/rôle.

Soit $subroles(f)$ l'ensemble des sous-ensembles des rôles de la frame $f$ tel que la taille de chaque sous-ensemble soit égal au nombre d'arguments candidats. On a :

$$subroles(f) = \begin{cases} \{\mathcal{R} \subset roles(f) / |\mathcal{R}| = |\mathcal{A}|\} & si\ |roles(f)| \geq |\mathcal{A}| \\ \{roles(f)\} & sinon \end{cases}$$

$$\forall \mathcal{R} \in subroles(f), \forall b \in \{b\ /\ \forall r \in \mathcal{R}, \exists! a \in \mathcal{A}, b(a) = r\},$$
$$scoreBij(b) = \frac{1}{|\{a \in \mathcal{A} / \exists r = b(a)\}|} \sum_{a \in \{a \in \mathcal{A} / \exists r = b(a)\}} scoreAsso(a, r)$$

La meilleur bijection est conservée et donne son score à la structure candidate, ce qui nous donne :

$$\forall c \in \mathcal{C}, scoreCand(c) = confiance(p). \max_{\mathcal{R} \in subroles(f), b \in \mathcal{R}} scoreBij(b)$$

avec confiance(p) le score obtenu par l'unité lexicale lors de la traduction de FrameNet.

#### 6.1.1.4 Validation des structures candidates

Pour chaque prédicat, la structure candidate de score le plus élevé donne la frame correspondant à l'usage du prédicat dans la phrase.

$$structure\ validée = argmax_{c \in \mathcal{C}} scoreCand(c)$$

Cependant, il existe deux cas dans lesquels les termes reconnus comme prédicats lors de la phase de détection ne le sont pas réellement. Le premier cas se produit lorsque le terme est effectivement un déclencheur de la frame reconnue, mais qu'il ne tient pas un rôle de prédicat dans le contexte de la phrase. Ce n'est pas un problème de désambiguïsation mais simplement de réalisation. Par exemple, dans la phrase *La perte d'emploi est un facteur de précarité.*, le terme *emploi* sera reconnu comme déclencheur de la frame *Being_employed*. Cependant, dans ce contexte-là, aucun des rôles principaux *Employee, Employer, Field, Place_of_employment, Position, Task* n'est réalisé. Le deuxième cas se produit lorsque la frame validée ne correspond pas au sens du prédicat dans la phrase. Ceci peut arriver dans le cas où la frame correcte n'existe pas dans la ressource, ou bien lorsque l'unité lexicale correspondant au prédicat n'est pas présente dans la frame correcte, ou encore lorsque le procédé de choix de la meilleure frame a échoué.

Dans chacun de ces cas de fausse détection, le score de la bijection retenu est généralement faible, voire nul. Nous fixons donc empiriquement un seuil en dessous duquel la frame est invalidée.

#### 6.1.1.5 Recouvrement des syntagmes arguments et annotation

À partir des arguments détectés sous forme de terme, nous procédons enfin à la détection du syntagme argument auquel chaque terme argument appartient. Ceci est effectué en suivant les relations syntaxiques gouvernées par le terme argument. Chaque terme ainsi trouvé est retenu comme appartenant au syntagme à condition qu'il ne chevauche pas un autre argument. On réitère de façon récursive en suivant les relations syntaxiques gouvernées par les termes ainsi retenus, en maintenant la condition de non-chevauchement.

Le prédicat et les syntagmes correspondants sont alors annotés en fonction des frames retenues et de leur structure associée. Notons que notre système permet la présence de plusieurs prédicats dans une même phrase.

### 6.1.2 Score d'association entre un argument candidat et un rôle

Nous définissons le score d'association entre un argument candidat et un rôle comme le produit du poids issu de la distance syntaxique et de deux scores distincts décrits respectivement dans les deux sous-sections suivantes. Nous ne traitons dans cette première étude que les rôles considérés comme principaux (*Core*) par FrameNet.

Ne disposant pas de corpus d'apprentissage sur la langue française, nous souhaitons néanmoins disposer de données sur les éléments lexicaux pouvant remplir les rôles sémantiques. Pour cela, nous utilisons le corpus anglais annoté et extrayons la tête de chaque syntagme annoté par un rôle. Nous associons ensuite l'ensemble des traductions de cette tête de syntagme, au rôle et à la frame correspondante et conservons ces listes. Les traductions sont données par nos dictionnaires bilingues EuRADic et Wiktionnaire (déjà utilisés et décrits en 3.1 et 5.1.1). Nous conservons uniquement les têtes de syntagmes d'une part car nous supposons que ce sont les termes portant la plus grande partie de l'information sémantique des syntagmes, et d'autre part car nos espaces sémantiques ont été construits à partir de relations de dépendance gouvernant des têtes de syntagme. Nous disposons ainsi pour chaque rôle de chaque frame d'une liste de lexèmes français correspondants.

Du fait de la polysémie potentielle des lexèmes source, ces ensembles de traductions sont parfois bruités mais nous supposons que la quantité de données générées sera suffisante pour que le bruit ne perturbe pas excessivement nos résultats.

#### 6.1.2.1 Projection lexicale et similarités distributionnelles

Nous cherchons à définir un score d'association entre un argument $a$ et un rôle $r$, permettant de déterminer si l'argument $a$ correspond bien au rôle $r$ de la frame $f$ dans la phrase analysée. Nous supposons que si le rôle $r$ a déjà été rempli par un certain nombre d'arguments dans le corpus FrameNet, l'argument $a$ correspondant au rôle $r$ est sémantiquement proche des arguments correspondants du corpus. Le système cherche donc à maximiser la similarité sémantique de notre argument $a$ avec les têtes des syntagmes annotés correspondant au rôle $r$. La similarité sémantique entre l'argument $a$ et le rôle $r$ est donnée par la moyenne des similarités sémantiques entre l'argument $a$ et chacun des lexèmes $l$ du rôle $r$, c'est-à-dire par le cosinus approximé des représentations de $a$ et de $l$. L'approximation est nécessaire pour pouvoir effectuer cette analyse en temps réel. Concernant les détails techniques de cette approximation, nous rappelons au lecteur que la présentation de l'approximation figure à la section 2.1.2.2.

Si la liste de lexèmes correspondant au rôle $r$ de la frame $f$ contient plus d'un certain nombre d'éléments (fixé empiriquement à 5 dans notre étude), le contexte est jugé suffisant et la proximité sémantique est calculée avec cette liste. Sinon, la liste est étendue aux listes de lexèmes réalisant le même rôle $r$ mais pour toutes les frames. Dans ce cas-là, les lexèmes correspondant à la frame $f$ ont une pondération double. Nous notons $lexemes(r)$ cette liste.

Enfin, ces similarités sémantiques sont calculées sur l'ensemble des relations syntaxiques significatives liant l'argument $a$ aux autres mots de la phrase. En effet, cela permet de réduire l'effet de polysémie, puisque nous avons montré que certains sens de mots s'emploient plus fréquemment dans certaines positions syntaxiques que d'autres (cf. 3.2.2). Cet ensemble est noté $relations(a)$.

Ainsi la première définition du score d'association que nous proposons est la suivante :

$$scoreAsso_1(a,r) = \frac{1}{|relations(a)|}\frac{1}{|lexemes(r)|}\sum_{rel\in relations(a)}\sum_{l\in lexemes(r)} cosinus_{rel}(a,l)$$

où $cosinus_{rel}(a,l)$ est le cosinus approximé des représentations de $a$ et $l$ dans l'espace sémantique correspondant à la relation $rel$.

Afin de ne pas disqualifier les termes ne figurant pas dans nos espaces sémantiques, nous leur attribuons des scores par défaut. Les pronoms ne possédant que peu de contenu sémantique propre mais étant très fortement susceptibles de réaliser un rôle lorsqu'ils apparaissent en relation avec un prédicat, nous leur attribuons un score de 0,3. Les noms propres ne figurent pas dans nos espaces mais ils sont également fortement candidat à être argument d'un prédicat, nous leur attribuons un score de 0,2. Enfin, les termes inconnus peuvent aussi réaliser des rôles, mais nous leur accordons moins de confiance, nous leur attribuons donc un score de 0,1. Le choix de ces scores par défaut est cependant arbitraire et justifierait un apprentissage.

Ce score suffirait à résoudre l'attribution des rôles pour la phrase très simple suivante :

*La souris mange le fromage.* (1)

En effet, la liste des traductions associée au rôle *Ingestor* et celle du rôle *Ingestibles* sont présentées dans le tableau 6.1. En calculant la similarité cosinus entre l'argument *souris* et les deux listes correspondant aux rôles, on s'aperçoit que le mot *souris* est plus proche de la liste de mots représentant le rôle *Ingestor*, grâce notamment aux éléments animés de cette liste puisqu'ils partagent en corpus un grand nombre de contextes verbaux de type sujet.

D'autre part, le mot *fromage* est plus proche des éléments correspondant au rôle *Ingestibles* tant grâce aux éléments que l'on peut effectivement *manger*, mais aussi grâce à tous les éléments qui partagent des contextes verbaux complément d'objet communs (on peut par exemple *consommer, avaler, acheter, jeter, mettre, utiliser, prendre* ou encore *donner une bière, un café, une glace, de la nourriture, un toast, du vin, (...)*).

Ces scores permettent de déterminer quelle est la bijection optimale entre arguments et rôles. Le fait de n'évaluer que les bijections implique qu'un rôle ne peut être attribué à plusieurs arguments, et vice-versa. Ainsi, les scores d'associations présentés dans le tableau 6.2 permettent sans hésitation de déterminer quel est l'argument correspondant au rôle*Ingestor* et quel est l'argument remplissant le rôle *Ingestibles*. En effet, la solution optimisant la somme des scores consiste à choisir la résolution {*Ingestor : La souris, Ingestibles : le fromage*}. Bien que la différence des scores ne soit pas significative entre les résolutions {*Ingestor : le fromage*} et {*Ingestibles : le fromage*} ou entre les résolutions {*Ingestibles : le fromage*} et {*Ingestibles : La souris*}, le fait que la résolution {*Ingestor : La souris*} ait un score beaucoup plus élevé permet par élimination d'attribuer le rôle de *Ingestibles* à l'argument *le fromage*.

| | |
|---|---|
| Ingestor | *aspic, bande, chambre, élève, enfance, enfant, espèce, famille, femelle, gens, gonzesse, greluche, habitants, île, jeune, jumelle, maison, mâle, marijuana, meuf, mite, môme, nana, nation, nénette, oiseau, parents, patron, pépée, personnes, petit, petite, peuplade, peuple, pièce, poiscaille, poisson, population, poule, poulette, race, souris, type, volaille, (...)* |
| Ingestibles | *alcool, aliment, apéritif, aspic, baille, ballon, baromètre, barre, biberon, bibine, bière, bocal, boisson, bouchée, bouffe, bourbon, bout, bouteille, brandy, breuvage, café, canette, canon, casse-croûte, champagne, chère, chopine, cognac, consommation, couleur, coupe, croûte, cru, cruche, cruchon, cuisine, cuvette, de, découpure, déjeuner, digestif, dîner, eau, en, engrais, essence, farine, fiole, flacon, flotte, fragment, frichti, glace, gobelet, gobeleterie, godet, gorgée, goutte, infusion, jus, kawa, la, lentille, loupe, manger, menu, miroir, moiré, morceau, nourriture, objets, outre, pain, passage, pôtée, pâture, petit, petit-déjeuner, pièce, pinard, plein, pot, repas, scotch, sirop, suc, tasse, thé, timbale, tisane, toast, verre, verrerie, vin, vitrerie, vitrine, whisky, (...)* |

TABLE 6.1 – Liste des traductions les plus fréquentes des têtes de syntagmes réalisant les rôles *Ingestor* et *Ingestibles*

| | SUJ : **La souris** | OBJ : **le fromage** |
|---|---|---|
| Ingestor | **0,66** | 0,55 |
| Ingestibles | 0,55 | **0,56** |

TABLE 6.2 – Score d'association lexicale pour la phrase (1)

En revanche, considérons la phrase suivante dont les deux arguments sont d'un type sémantique animé :

*Le chat mange la souris.* (2)

Les scores d'associations présentés dans le tableau 6.3 montrent que même si la décision est prise à bon escient, le score optimal est moins élevé et plus proche de la solution inverse. Ici, *chat* en tant que sujet est plus proche (par le cosinus) de la liste des têtes de syntagmes remplissant le rôle de *Ingestor*, tandis que *souris* en tant que complément d'objet est plus proche de la liste des têtes de syntagmes remplissant le rôle de *Ingestibles*.

| | SUJ : **Le chat** | OBJ : **la souris** |
|---|---|---|
| Ingestor | **0,70** | 0,62 |
| Ingestibles | 0,56 | **0,57** |

TABLE 6.3 – Score d'association lexicale pour la phrase (2)

Enfin, nous considérons la phrase suivante de sens opposé :

*La souris mange le chat.* (3)

dont les scores figurent dans le tableau 6.4. L'identification des rôles sémantiques est encore une fois correctement résolue par la combinaison {*Ingestor : La souris, Ingestibles : le chat*} mais la différence avec la solution inverse est encore plus faible. Ici, *souris* en tant que sujet est plus

proche de la liste des têtes de syntagmes remplissant le rôle de *Ingestor*. En ce qui concerne le second argument, *chat* en tant que complément d'objet est plus proche de du rôle *Ingestibles* que ne l'est *souris*, en revanche, il est plus proche du rôle *Ingestor* que du rôle *Ingestibles*. Cependant, l'optimisation globale du score permet une identification correcte.

| manger.v | SUJ : **La souris** | OBJ : **le chat** |
|---|---|---|
| Ingestor | **0,66** | 0,67 |
| Ingestibles | 0,55 | **0,59** |

TABLE 6.4 – Score d'association lexicale pour la phrase (3)

Supposons maintenant que l'on doive analyser la phrase suivante :

*La souris nourrit le chat.* (4)

D'après le tableau 6.5, cette phrase est résolue avec les mêmes scores que la phrase (3), par la combinaison {*Ingestor : La souris, Ingestibles : le chat*}. Dans ce cas-là, cela est incorrect. En effet, le score lexical $ScoreAsso1$ ne tient pas compte du prédicat, mais observe seulement les plus proches voisins des arguments en fonction de leur position syntaxique.

| nourrir.v | SUJ : **La souris** | OBJ : **le chat** |
|---|---|---|
| Ingestor | **0,66** | 0,67 |
| Ingestibles | 0,55 | **0,59** |

TABLE 6.5 – Score d'association lexicale pour la phrase (4)

La considération de ces cas plus ambigus nous mène à la définition du deuxième élément de notre score d'association à la sous-section suivante.

#### 6.1.2.2 Projection syntaxique et associations distributionnelles

L'argument $a$ que nous cherchons à attribuer au rôle $r$ disposant au moins d'une annotation précisant ses relations syntaxiques avec le reste de la phrase, il devient très intéressant d'exploiter ce contexte syntaxique. Nous connaissons en effet les différentes lexicalisations du rôle $r$ (données par nos listes de traduction), et cela nous permet d'observer si ce rôle est souvent lexicalisé dans un tel contexte syntaxique. Si c'est le cas, on peut considérer que l'association entre l'argument $a$ et le rôle $r$ est forte. Cette quantification de la force d'association est donnée par le score suivant :

$$scoreAsso_2(a,r) = \frac{1}{|lexemes(r)|} \sum_{l \in lexemes(r)} \frac{1}{|relations(a)|} \sum_{rel \in relations(a)} PMI_{rel}(contexte_{rel}(a), l)$$

avec $relations(a)$ l'ensemble des relations gouvernant l'argument $a$, $PMI_{rel}(contexte, lexeme)$ l'information mutuelle spécifique entre *contexte* et *lexeme* contenue dans la matrice de l'espace sémantique correspondant à la relation syntaxique $rel$, et $contexte_{rel}(l)$ le contexte syntaxique donné par la relation syntaxique $rel$ gouvernant le lexème $l$ dans la phrase.

Notons que les espaces sémantiques ignorés dans les autres parties de nos travaux car considérés comme non significatifs ont tout à fait leur place ici. Par exemple, l'espace issu de la relation syntaxique *préposition-substantif* peut aider à résoudre de nombreux cas d'association en particulier pour les rôles *Non-Core* (*Location*, *Time*, *Manner*,...). Cependant, pour une raison que nous ignorons, cet espace n'a pas été constitué lors de la construction des espaces sémantiques que nous utilisons. L'intégration de cet espace reste donc à l'état de perspective.

Enfin, les relations de l'analyseur syntaxique LIMA gouvernant spécifiquement les pronoms sont mises en correspondance avec les relations nominales correspondantes. Par exemple, pour les relations *complément d'objet direct pré-verbal* et *complément d'objet indirect pré-verbal*, on calcule les scores d'association respectivement avec les espaces standard *complément d'objet direct* et *complément d'objet indirect.*

Ce score permet de résoudre le problème soulevé plus haut des phrases (2), (3) et (4). Les tableaux 6.6, 6.7 et 6.8 présentent respectivement les scores de chacune de ces phrases.

Pour calculer ce score, on réutilise les listes présentées dans le tableau 6.1 et on calcule l'information mutuelle liant chacun des termes dans les matrices syntaxiques correspondant à la relation qui lient les arguments au prédicat. Par exemple, pour la phrase (2), le système calcule l'information mutuelle entre chacun des mots *aspic, bande, chambre, élève, enfance, enfant, espèce, famille, femelle, gens, gonzesse, (...)* représentant le rôle *Ingestor* et le prédicat *manger* dans la matrice construite avec la relation syntaxique *sujet du verbe* pour donner un score d'association syntaxique de 0,69. L'opération est répétée avec chacun des rôles et chacune des relations syntaxiques.

| | **SUJ** : Le chat | **OBJ** : la souris |
|---|---|---|
| Ingestor | **0,69** | 0,69 |
| Ingestibles | 0,42 | **0,52** |

TABLE 6.6 – Score d'association syntaxique de la phrase (2)

| manger.v | **SUJ** : La souris | **OBJ** : le chat |
|---|---|---|
| Ingestor | **0,69** | 0,69 |
| Ingestibles | 0,42 | **0,52** |

TABLE 6.7 – Score d'association syntaxique de la phrase (3)

On voit dans les tableaux 6.6 et 6.7 que les scores des positions syntaxiques *sujet* et *objet* sont identiques dans chacune des deux phrases. En effet, le score syntaxique $ScoreAsso2$ ne tient pas compte du lexème qui tient le rôle de l'argument mais calcule la force d'association entre le prédicat et la liste des têtes de syntagmes remplissant le rôle candidat dans la position syntaxique donnée.

Enfin, dans le cas du prédicat *nourrir.v* le score n'est pas très informatif. Cela est dû au fait que le verbe *nourrir.v* s'emploie très régulièrement avec une structure de sous-catégorisation différente,

| nourrir.v | **SUJ** : La souris | **OBJ** : le chat |
|---|---|---|
| Ingestor | 0,62 | **0,65** |
| Ingestibles | **0,31** | 0,28 |

TABLE 6.8 – Score d'association syntaxique de la phrase (4)

comme c'est le cas dans *La mère nourrit son enfant*. Pour cette raison, la force d'association entre le prédicat et toutes les têtes de syntagmes de type animé et en position de sujet sont élevées. De ce fait, le rôle *Ingestor*, contenant un grand nombre de têtes de syntagme de type sémantique animé, obtient un score d'association plus fort que souhaité en position de sujet. Ici, le score optimal donne cependant la bonne résolution.

Dans ces cas simples, les relations syntaxiques gouvernant les arguments de ces deux phrases suffisent à résoudre l'attribution des rôles.

### 6.1.3 Aspects innovants de la méthode

Parmi les récentes approches semi-supervisées, cette méthode est complètement innovante. D'une part, nous ne connaissons pas de système ayant tenté une annotation non supervisée en rôles de types FrameNet. D'autre part, à notre connaissance, aucune méthode n'a jamais exploité l'information distributionnelle de la façon dont nous le faisons.

Les travaux auxquels nous pourrions nous comparer sont d'une part ceux de [Abend *et al.* 2009], bien que ceux-ci concernent uniquement la tâche de détection des arguments. Ceux-ci exploitent une information mutuelle spécifique issue de cooccurrence de fenêtre, qu'ils calculent entre les prédicats et les candidats à devenir argument. Le seul but de cette utilisation est de filtrer les candidats n'ayant que peu de lien sémantique avec le prédicat.

Dans notre méthode, d'une part nous ne calculons pas la même information mutuelle : les nôtres (utilisées dans le score d'association syntaxique) sont des informations mutuelles spécifiques à des relations syntaxiques données ; cependant, la différence la plus grande est que nous ne la calculons pas entre un prédicat et un argument mais entre un prédicat et une liste d'arguments potentiels. Ceci permet de déterminer si la relation syntaxique utilisée avec le prédicat donné est une bonne position syntaxique pour un argument.

Cela permet de se rapprocher d'une approche supervisée tout en apportant une généralisation complémentaire. En effet, chaque fois que la force d'association syntaxique entre un prédicat et un argument supposé est suffisamment forte, elle permet de simuler une instance d'argument doté d'un trait syntaxique et associé au prédicat donné. Les équivalents anglais de ces instances n'apparaissent pas nécessairement dans le corpus anglais d'origine. Dans ces premiers travaux, nous n'avons pas développé l'aspect apprentissage supervisé mais nous souhaitons développer celui-ci dans nos perspectives. En revanche, nous exploitons ici cette généralisation par l'utilisation de nos scores.

Les seconds travaux auxquels nous pouvons nous comparer sont ceux de [Swier & Stevenson 2004]. En effet, ceux-ci proposent une méthode semi-supervisée par *bootstrapping*. L'idée est de calculer pour chaque argument candidat les probabilités conditionnelles $P(role|predicat, argument, syntaxe)$. Ceci n'étant pas possible du fait du peu de données produites à l'initialisation du bootstrapping, leur méthode propose plusieurs généralisations du calcul de modèles de probabilités par relâchement de diverses contraintes. Ils calculent ainsi notamment les probabilités $P(role|syntaxe, classe\_de\_verbes)$ et $P(role|verbe, classe\_de\_noms)$.

Leur généralisation des probabilités conditionnelles aux classes de verbes issus de VerbNet est intrinsèque à la structure de VerbNet. En effet, dans VerbNet chaque classe de verbes est définie par une même structure sémantico-syntaxique. Dans notre méthode, le score d'association syntaxique produit une généralisation similaire pour les classes d'unités lexicales que sont les frames, à la grande différence que les cadres de sous-catégorisation des unités lexicales ne sont pas nécessairement identiques à l'intérieur d'une même frame. Par ailleurs, nous ne calculons pas de probabilité dans notre méthode mais nous nous contentons de prendre en compte cette généralisation par l'usage du score d'association.

De la même façon, on peut également faire le parallèle entre leur généralisation des probabilités conditionnelles à l'aide de classes de noms issues des catégories les plus hautes de WordNet, et notre score d'association lexicale. Les classes de noms sous-jacentes à notre score n'ont certes pas la même granularité puisqu'il s'agit de classes correspondant directement aux rôles des frames détectées, mais l'idée générale est la même. Il s'agit d'étendre la connaissance du système pour qu'il puisse prendre des décisions en se référant non plus aux rares instances du lemme de l'argument candidat rencontrées dans le corpus annoté, mais également aux instances de termes de la même classe. Une fois de plus, notre système n'emploie pas une approche d'apprentissage supervisé et ne calcule donc pas de probabilités à l'aide de ces classes. En revanche, le score d'association lexicale permet de rapprocher sémantiquement l'argument candidat de la classe correspondant le plus probablement à son rôle.

Notre approche étant ainsi définie, nous pouvons à présent passer à l'évaluation de notre système.

## 6.2 Évaluation

Nous souhaitons procéder d'une part à l'évaluation de la projection lexicale, d'autre part à l'évaluation de la projection syntaxique et enfin à la combinaison de celles-ci, afin de valider ou non les stratégies proposées.

### 6.2.1 Protocole expérimental

À l'heure actuelle, seul le corpus produit par [Padó & Pitel 2007] dispose d'annotations en rôles sémantiques pour la langue française. Celui-ci contient un ensemble de 1076 phrases extraites du corpus Europarl ([Koehn 2005]) et annotées à l'aide des rôles de FrameNet 1.1.

Nous proposons donc de procéder à deux phases dans notre protocole d'évaluation. La première utilise les mesures standards proposées par les campagnes Senseval et compare les sorties de notre système au corpus produit par [Padó & Pitel 2007]. Nous considérerons ce corpus comme une pseudo-vérité-terrain car il est lui-même issu d'une analyse automatique.

Nous utilisons le programme d'évaluation fournit par la campagne SemEval 2007 ([Baker *et al.* 2007]). Celui-ci calcule la précision, le rappel et la F-mesure sur la tâche de reconnaissance et classification à la fois des frames et des rôles. Nous éliminons le score associé aux rôles *Non-core* puisque nous avons restreint cette étude aux rôles *Core*.

La deuxième phase de ce protocole vise à étudier les cas non traités par nos pseudo-vérités-terrains et à valider ou non les sorties de notre analyseur de façon manuelle.

### 6.2.2 Résultats

Nous rapportons et commentons dans un premier temps les résultats de l'évaluation automatique, puis procédons à l'analyse des différences entre les sorties de notre système et le corpus de [Padó & Pitel 2007].

### 6.2.3 Évaluation automatique

L'évaluation automatique des frames et roles identifiés par notre système en comparaison avec le corpus de [Padó & Pitel 2007] donne les résultats fournit dans le tableau 6.9. Les mesures sont calculées de la même façon que ce qui a été fait dans SemEval 2007, c'est-à-dire en considérant l'annotation des prédicats et des rôles avec un poids identique, et par conséquent en comptant nécessairement comme fausses les occurrences de rôles des prédicats eux-mêmes mal annotés.

| | Précision | Rappel | F-mesure |
|---|---|---|---|
| Score lexical | 4 | 12 | 5 |
| Score syntaxique | 5 | 5 | 6 |
| Combinaison des scores | 4 | 9 | 6 |

TABLE 6.9 – Mesures de pertinence en pourcentage

Rappelons que les F-mesures des systèmes supervisés ayant participé à la campagne SemEval 2007 varient entre 49 et 75% selon les corpus pour l'identification de la frame et entre 37 et 49% pour l'identification des rôles. Malgré le fait que notre système ne soit que faiblement supervisé, ces résultats semblent à première vue assez décevants.

La seule chose que nous pouvons inférer à partir de ces résultats automatiques est que le score syntaxique semble permettre une annotation plus précise que le score lexical. Ce dernier permet en revanche de rapporter plus de résultats et d'obtenir ainsi un meilleur rappel. La combinaison des scores n'améliore pas significativement les résultats mais ne les détériore pas non plus. Il faudrait revoir l'opérateur de combinaison afin d'utiliser mieux les propriétés de chaque score.

Nous allons maintenant observer les données par une analyse manuelle d'échantillons. Nous cherchons d'une part à savoir si les frames détectées par notre système et non annotées dans le corpus sont effectivement incorrectes ou non, et d'autres part quelles sont les causes d'échec de notre système.

### 6.2.4 Analyse manuelle

Nous analysons ainsi manuellement des échantillons de 10 éléments, couvrant les cas suivants :

**Cas n°1** Frames détectées par notre système mais absentes du corpus de référence (faux positifs)

**Cas n°2** Phrases contenant des frames détectées par notre système mais dont la phrase dans le corpus de référence n'est pas annotée du tout (ignorés)

**Cas n°3** Frames détectées dans le corpus de référence mais pas par notre système (faux négatifs)

**Cas n°4** Rôles annotés par notre système mais différent du corpus de référence (faux positifs)

**Cas n°5** Rôles manqués par notre système mais annotés dans le corpus de référence (faux négatifs)

Sauf mention contraire, les analyses sont données à partir des résultats issus du score combiné.

#### 6.2.4.1 Cas n°1

Parmi les dix frames détectées et identifiées par notre système et non présentes dans le corpus de référence, cinq frames ont été jugées correctement identifiées et une sixième acceptable même si l'annotation est moins précise que celle donnée par le corpus de référence. On a donc un taux légèrement supérieur à 50%. Parmi ces six frames valides cinq sont également obtenues à partir du score lexical seul, et trois sont obtenues à partir du score syntaxique seul. Cela conforte l'idée intuitive de la complémentarité des deux scores.

Concernant les frames erronées, la source majoritaire d'erreurs concerne des prédicats présents dans notre FrameNet français mais appartenant à des frames ne convenant pas pour l'utilisation dans les phrases concernées. C'est le cas par exemple du prédicat *décrire* dans la phrase *Quant à moi, je répète que, dans ce domaine, obliger le demandeur de brevet à décrire son invention est très important.* En effet, la seule frame dans laquelle ce prédicat apparaît dans notre FrameNet est la frame *Communicate_categorization* décrivant une situation dans laquelle un orateur établit l'appartenance d'un item à une catégorie. Dans le cas précis de cette phrase, notre système a identifié le prédicat comme appartenant à cette frame, menant ainsi à ce type d'erreurs visiblement assez récurrent.

Une autre erreur provient d'un lien syntaxique manquant n'ayant pas permis le rattachement d'un argument candidat et ayant ainsi généré une erreur de désambiguïsation. C'est le cas dans la phrase suivante où seule le lemme *voie* a été relié par une relation syntaxique au prédicat *poursuivre* :

*Je pense que nous devrions nous efforcer de* ***poursuivre*** *sur cette voie dans les domaines de la poste, de l'énergie, des télécommunications et des chemins de fer ; domaines où règnent des conditions comparables.*

Le prédicat *poursuivre* n'a donc plus que pour argument candidat *sur cette voie*. Dans notre FrameNet, ce prédicat appartient à la frame *Seeking_to_achieve* correspondant au sens utilisé dans cette phrase. En revanche, il appartient aussi à la frame *Cotheme* faisant intervenir deux éléments mobiles se déplaçant l'un par rapport à l'autre. L'unique contexte utilisé pour le calcul du score faisant référence à une *voie*, le score lexical marque la différence et c'est la frame *Cotheme* qui est préférée générant ainsi une erreur.

Enfin, la frame moins précise que celle de la référence est issue de la non-détection de l'expression figée *mettre en danger* par notre analyseur linguistique, et de son absence dans notre FrameNet. En effet, le corpus a pour annotation *Endangering*, tandis que notre système se contente de la frame *Placing*.

#### 6.2.4.2 Cas n°2

Parmi les 10 frames détectées par notre système dans des phrases non analysées dans le corpus de référence, notre système détecte en moyenne 2,1 prédicats par phrases, parmi lesquels 1,8 sont effectivement des prédicats et 1,0 sont identifiés par la frame correcte. On retrouve notre taux approximatif de 50% de frames correctement annotées.

#### 6.2.4.3 Cas n°3

Dans le cas des frames présentes dans la référence et non détectées par notre système, les erreurs sont assez variées. Une des annotations n'était pas réellement une erreur mais une détection de la forme complète du prédicat comprenant son auxiliaire de temps, ce qui n'était pas le cas dans la référence.

On retrouve pour trois d'entre ces cas la problématique de la non-exhaustivité de notre FrameNet. En effet, dans ces trois cas, les paires frames-prédicats annotées dans la référence sont absentes de notre FrameNet.

Dans les autres cas, la paire prédicat-frame présente dans la référence est trouvée par notre système dans FrameNet, mais le score obtenue pour cette structure candidate n'est pas suffisamment élevé. Soit il est inférieur au seuil de confiance, soit il est inférieur au score d'une autre structure candidate concernant le même prédicat mais une autre frame. Dans les deux cas, la frame n'est pas annotée. Nous invoquons différentes raisons pour lesquelles ces scores restent trop faibles :

– lien manquant dans l'analyse syntaxique ;

– mauvaise détection de l'étiquette morphosyntaxique ;
– distance trop grande d'un argument fortement désambiguïsateur ;
– liste vide de lexèmes pouvant remplir un rôle donné.

#### 6.2.4.4 Cas n°4

Concernant les 10 rôles détectés par notre système et absents de la référence, huit concernent des frames soit présentes dans la référence (correctement détectées par l'évaluateur), soit des frames que nous avons jugées correctes même si elles sont absentes ou différentes de la référence.

On a donc 20% de frames invalides pour lesquelles les rôles associés ne sont pas ou mal applicables.

Parmi les frames valides, les différents cas se répartissent de la façon suivante : un des rôles est jugé acceptable même si différent de la référence et deux sont issues de la présence d'une liste de lexèmes vide pour un des rôles associés.

Enfin deux rôles concernent des frames non présentes dans la référence. Dans ces deux cas, soit le rôle qu'il faudrait utiliser pour l'annotation est *Non-core* et notre système ne le connaît pas, soit le rôle qu'il faudrait utiliser nous a semblé mériter d'exister mais il n'est pas présent dans FrameNet.

Nous donnons cette dernière indication à titre d'information. Cependant, combien même cette identification aurait été correcte, les rôles associés à des frames non validées par la référence sont de toutes façons comptés faux car les rôles sont spécifiques à chaque frame.

#### 6.2.4.5 Cas n°5

Enfin, parmi les 10 rôles manqués par notre système figurent trois rôles associés à des prédicats que notre système a manqué, ainsi qu'un rôle associé à un prédicat que notre système a détecté mais pour lequel il a attribué une mauvaise frame (lien syntaxique manquant, absence de la paire frame-prédicat dans notre FrameNet, liste vide pour les rôles).

Les autres frames concernées sont égales ou équivalentes aux frames de la référence (60%), voir même plus exacte pour un des cas. Les erreurs d'identification de rôles restant applicables parmi ces cas-là sont au nombre de trois et proviennent soit de liens syntaxiques manquants, soit du design même de notre système qui n'est pas suffisamment robuste pour certains cas d'ambiguïté.

### 6.2.5 Synthèse

Notre travail sur l'annotation en rôles sémantiques est une première étape vers l'exploitation de l'information distributionnelle syntaxique au cœur de l'annotation elle-même. Nous convenons que la méthode manque de maturité et que certains choix nécessiteraient d'être faits par un apprentissage. La combinaison des scores reste notamment un élément à revoir.

Nous pouvons néanmoins esquisser une première analyse, car l'introspection manuelle d'annotations erronées lors de l'analyse des données nous suggère plusieurs facteurs d'erreurs récurrents

mettant en cause différentes informations que notre système considère comme disponibles.

En effet, une mauvaise annotation syntaxique perturbe d'une part le choix des arguments candidats, d'autre part les scores d'association qui leur sont attribués, et peut même ainsi influencer le choix de la frame pour un prédicat donné. L'analyse syntaxique reste donc un élément donc l'annotation en rôles sémantiques est très fortement dépendante. Rappelons aussi qu'à l'inverse, l'analyse syntaxique est parfois également tributaire de l'analyse sémantique, notamment dans les cas d'ambiguïté du rattachement des syntagmes prépositionnels. Il pourra donc être intéressant d'observer dans une phase ultérieure, s'il existe des cas où les annotations sémantiques sont correctes alors que l'annotation syntaxique ne l'est pas.

Deux autres sources d'erreurs majoritaires proviennent des données de FrameNet. En effet, nous nous sommes souvent trouvés confrontés à un prédicat dont la frame pour la phrase donnée ne lui est pas associée dans FrameNet. Il serait donc intéressant de continuer à enrichir FrameNet, particulièrement pour les prédicats verbaux. Par ailleurs, nous nous retrouvons malgré tout confrontés à des problèmes de manque de corpus annoté. En effet, certains rôles sont définis dans FrameNet mais n'ont aucune annotation correspondante en corpus, et il n'y a aucun moyen de connaître le type de lexicalisation de ces rôles.

Enfin, la revue des différents travaux de la littérature en analyse de rôles semi-supervisée montre que cette tâche n'a encore que très peu été abordée sous cette perspective. Ceci est cependant une tendance de plus en plus étudiée, essentiellement pour appréhender d'autres langues que l'anglais. Les approches que nous connaissons [Swier & Stevenson 2004] et [Abend *et al.* 2009] utilisent respectivement le référentiel VerbNet et PropBank. La tâche dédiée à FrameNet est plus complexe du fait de la désambiguïsation nécessaire et du nombre de rôles important. Les résultats ne sont donc pas comparables. En effet, ces systèmes semi-supervisés utilisés avec VerbNet et PropBank obtiennent des mesures supérieures aux meilleurs systèmes supervisés utilisant FrameNet.

## 6.3 Intérêt de l'annotation de rôles sémantiques en Recherche d'Information

De cette représentation du texte en rôles sémantiques, deux points intéressants peuvent être soulignés pour la recherche d'information. D'une part, les annotations en rôles sémantiques peuvent s'avérer utile pour répondre aux problèmes posés par la recherche par mots-clés. Nous avons déjà présenté à la section 2.5.2 l'intérêt de l'extension de requête, tant pour élargir la recherche que pour la restreindre au cadre qui intéresse l'utilisateur (désambiguïsation). Une annotation en rôles sémantiques permettrait de retrouver les prédicats des exemples que nous avions mis en avant dès le premier chapitre et que nous reprenons dans la première sous-section. Nous étudions dans un premier temps l'intérêt d'une analyse effectuée uniquement sur la requête, et dans un deuxième temps l'intérêt d'une analyse effectuée également lors de l'indexation.

D'autre part, l'annotation en rôle sémantique peut jouer un rôle prépondérant dans les systèmes de Q/R. Nous proposons une illustration de l'usage des rôles sémantiques, en particulier lors de la recherche des passages pertinents et lors de l'extraction de la réponse.

### 6.3.1 Analyse de la requête sans indexation dédiée

Supposons que les requêtes *école de voile du port* et *port du voile à l'école* soit envoyées au moteur de recherche. Notre analyse en rôles sémantiques fournirait les résultats suivants :

$$[\text{école}]_{Local\_by\_use}\ \underline{\text{de voile}}_{Use}\ \underline{\text{du port}}_{Relative\_locale}$$
$$[\text{port}]_{Wearing}\ \underline{\text{du voile}}_{Clothing}\ \underline{\text{à l'école}}_{Place}$$

Sans même effectuer d'annotation sémantique de rôles lors de l'indexation, la simple analyse de la requête permettrait une recherche enrichie et non ambiguë. Par exemple, la simple détection de la frame *Local_by_use* dans la requête *école de voile du port* fournit une reformulation de la requête. En effet, en utilisant les différentes unités lexicales de la frame *Local_by_use*, on peut générer une requête comme *[centre/club]*$_{Local\_by\_use}$ *de voile du port*).

### 6.3.2 Analyse en rôles lors de l'indexation

Effectuer une telle annotation lors de la phase d'indexation des documents pourrait donner lieu à la mise en place de fonctionnalités supplémentaires.

#### 6.3.2.1 Contrôle de l'extension de requêtes

Un système d'extension de requêtes pourrait être contrôlé par l'usage de cette analyse. En effet, la détection de la frame lors de l'analyse de la requête permettrait un contrôle du sens des extensions de requête. Par exemple, si un système d'extension propose d'étendre *voile* à (*voile OR burqa*), $\underline{burqa}_{Use}$ n'a que très peu de chances de retourner des résultats, contrairement à $\underline{burqa}_{Clothing}$.

#### 6.3.2.2 Recherche sémantique

En outre, avec un tel type d'indexation, le moteur de recherche pourrait être interrogé de façon plus large. En proposant à l'utilisateur une interface graphique liée à un langage de requête un peu plus complexe, le moteur pourrait recevoir une requête équivalente à :

$$[\text{école}]_{Local\_by\_use}\ \underline{*}_{Use}\ \underline{\text{du port}}_{Relative\_locale}$$

Dans cette requête l'astérisque signifie que l'utilisateur ne souhaite pas spécifier la lexicalisation du rôle recherché. L'indexation par rôles permettrait de retourner à l'utilisateur des résultats sur toutes les écoles situées dans un port et concernant n'importe quelle activité ayant été reconnue comme *Use* d'un élément *Local_by_use* lors de l'indexation.

### 6.3.3 Indexation à l'aide des rôles pour les systèmes de Q/R

Ce dernier type de cas d'usage nous mène aux deuxième aspect que nous voulions souligner, à savoir, l'intérêt pour les systèmes de recherche de réponse aux questions posées en langage naturel. Une analyse en rôles sémantiques d'une question en langage naturel permet de typer la question avec une granularité plus précise que la plupart des systèmes sans SRL ne le font. Si les documents sont également indexés avec toute l'information disponible concernant les rôles, le système peut alors effectuer des traitements plus complexes.

Prenons pour exemple la question suivante :

$$\underline{\text{Qui}}_{Cognizer}\ [\text{a inventé}]_{Achieving\_first}\ \underline{\text{l'imprimerie}}_{New\_idea}\ ?$$

La reconnaissance du mot interrogatif *qui* remplissant le rôle de *Cognizer* de la frame *Achieving_first* permet d'interroger le système sous la forme évoquée précédemment. La requête est donc reformulée d'une part en remplaçant le pronom interrogatif par sa forme anonyme *_*Cognizer* afin de détecter tout type de réponse annotée par ce même rôle. D'autre part, le prédicat est également reformulé par sa forme anonyme afin d'étendre la recherche aux autres unités lexicales de la frame, ce qui est semblable à effectuer une extension de requêtes aux synonymes du prédicat et même plus riche car d'autres catégories morphosyntaxiques peuvent être détectées (passage d'un verbe à un nom, à un adjectif, et vice-versa). La nouvelle requête serait alors de la forme :

$$\underline{*}_{Cognizer}\ [*]_{Achieving\_first}\ \underline{\text{l'imprimerie}}_{New\_idea}\ ?$$

D'une part, cette requête permet au système de récupérer le passage suivant, contenant deux fois les frames, rôles et termes détectés dans la requête :

$\underline{\text{Si les Chinois}}_{Cognizer}$ [sont] bien [à l'origine de]$_{Achieving\_first}$ $\underline{\text{l'imprimerie}}_{New\_idea}$, pour un grand nombre d'historiens, $\underline{\text{l'imprimerie moderne}}_{New\_idea}$ [reste le fait de]$_{Achieving\_first}$ $\underline{\text{l'Allemand Johan Gutenberg qui, vers 1440, fond des caractères mobiles pouvant être réutilisés indéfiniment}}_{Cognizer}$.[1]

D'autre part, grâce aux analyses de la question et du passage cité à l'aide des rôles que nous avons identifié ici (*Cognizer* et *New_idea*), le module d'extraction de la réponse peut assez directement donner pour réponse *les Chinois* et pourrait même préciser que *l'imprimerie moderne [a été inventée par] l'Allemand Johan Gutenberg (...)*.

### 6.3.4 Perspectives

Afin de valider ces hypothèses et dans l'attente de premiers retours quant à l'intérêt de cette navigation, il serait très intéressant de mettre en œuvre d'une part un prototype permettant à

1. http://cerig.efpg.inpg.fr/dossier/impression-numerique/page01.htm

l'utilisateur de naviguer très précisément dans les données à l'aide de la représentation en rôles sémantiques.

D'autre part, intégrer notre module d'annotation de rôles à un système de Q/R existant permettra d'analyser formellement le gain de performance obtenu avec un système de SRL.

## 6.4 Conclusions

Nous avons dans ce chapitre proposé une approche complètement novatrice en analyse de rôles sémantiques. Nous introduisons dans notre algorithme l'usage d'informations syntaxiques distributionnelles. Ce type d'informations n'était jusque-là exploité que par les algorithmes supervisés, et ceci de façon implicite par l'exploitation du corpus d'apprentissage.

Notre approche permet de résoudre l'attribution de rôles pour des phrases simples de langue française. Pour des phrases plus complexes, notre analyse manuelle montre que notre système est très fortement dépendant à la fois de la ressource FrameNet utilisée et de l'analyseur en dépendances syntaxiques.

Nous avons également décrit comment un tel système pourra s'interfacer avec un système de recherche d'information, que ce soit pour des requêtes simples ou pour répondre à des questions posées en langue naturelle.

# Quatrième partie

# Conclusion

# Chapitre 7

# Bilan et perspectives

Nous concluons ce manuscrit par un rappel des différentes contributions de notre étude et une évocation des perspectives que nos travaux nous inspirent.

## 7.1 Contributions

Notre travail a porté d'une part sur la constitution automatique de ressources de langue française pour l'analyse sémantique de texte, d'autre part sur le développement de méthodes semi-supervisées à la fois pour la désambiguïsation lexicale et l'analyse en rôles sémantiques. Dans le cas de l'analyse en rôles sémantiques, c'est il s'agit d'un travail encore très initial mais qui introduit une idée encore non exploitée jusqu'à présent.

L'ensemble des différents modules proposés, implémentés et étudiés exploitent tous une ressource commune : un ensemble d'espaces distributionnels construits par [Grefenstette 2007] grâce à l'analyse en dépendances syntaxiques produite par LIMA ([Besançon & Chalendar (de) 2005]) d'un corpus de deux millions de documents.

Nous avons revu la pondération de ces espaces distributionnels à l'aide de l'information mutuelle, permettant ainsi aux termes peu fréquents d'être aussi bien représentés que les autres. Nous avons également appliqué l'algorithme de réduction de dimensions LSH de [Indyk & Motwani 1998] et [Charikar 2002]. Ceci nous a permis de manipuler nos données dans des temps raisonnables, que ce soit pour le calcul de la similarité cosinus ou encore pour la recherche rapide de plus proches voisins, pour laquelle il existe un algorithme spécifique aux signatures issues du hachage LSH.

### 7.1.1 Ressources

Nos contributions ont porté à la fois sur des ressources acquises de façon complètement automatique et sur des ressources construites automatiquement à partir de ressources préexistantes

constituées manuellement.

#### 7.1.1.1 Acquisition de ressources

Nous nous sommes intéressée à la section 2.2.2.4 de l'état de l'art aux différents algorithmes d'induction automatique de sens existants. Nous avons alors proposé au chapitre 3.2 deux méthodes inspirées de la littérature pour produire notre propre ensemble de clusters de mots représentant les sens des mots du vocabulaire.

L'originalité des algorithmes de clustering proposés consiste en la possibilité de prendre en compte différents espaces représentant les mêmes mots tout en tenant compte distinctement de l'information présente dans chaque espace. Ceci est rendu possible par l'utilisation d'une stratégie de votes ainsi que par l'étude de différentes possibilités de choix des noyaux.

De plus, nous avons souhaité appliquer ces algorithmes à des ensembles de plus proches voisins. Nous avons ainsi utilisé la méthode de recherche rapide de [Charikar 2002], que nous avons modifiée pour réduire sa complexité et améliorer sa pertinence.

Enfin, l'originalité des clusters produits réside quant à elle dans le choix des éléments à regrouper. En effet, nous proposons d'appliquer le clustering sur un ensemble constitué essentiellement de plus proches voisins et de cooccurrents du second ordre. Ces plus proches voisins ont été calculés indépendamment sur chacun des espaces correspondant à une relation syntaxique. Le fait d'être des cooccurrents syntaxiques leur confère une propriété particulière : ce sont des mots non seulement proches sémantiquement mais aussi syntaxiquement. Nous entendons par là qu'ils sont employés dans les mêmes contextes syntaxiques et par conséquent assez proches de la synonymie ou de l'antonymie.

Bien que cela n'ait pas été vérifié par une analyse automatique du fait de la complexité de sa mise en place, les clusters de sens obtenus semblent visuellement moins précis que ceux des méthodes originales. En effet, une inspection manuelle sur le mot *barrage* montre d'une part que certains clusters auraient pu être mieux regroupés, et d'autre part que les mots composant ces clusters sont moins descriptifs pour l'intuition humaine. Nous n'avons pas inspecté d'autres mots car les résultats ne sont pas fournis pour les deux algorithmes originaux.

Nos clusters de sens présentent néanmoins une discrimination de sens certaine, que nous n'avons pu observer que manuellement dans un premier temps. De plus, le fait de regrouper des plus proches voisins calculés sur des cooccurrences syntaxiques, ainsi que des cooccurrents syntaxiques du second ordre, confère à ces clusters de sens une propriété de similarité syntaxique. Nous entendons par similarité syntaxique le fait que ces mots sont employés dans des contextes syntaxiques proches de ceux du mot ambigu (par exemple *barrière* ou *écluse* pour le mot *barrage*). Il ne s'agit plus simplement de termes appartenant au même champ lexical (comme *tir* ou *rivière* pour le mot *barrage*). Cette propriété est essentielle pour l'algorithme de désambiguïsation lexicale que nous

proposons au chapitre 4. Cet algorithme a par ailleurs permis de valider la discrimination des sens générés.

#### 7.1.1.2 Traduction de ressources et enrichissement

Nous avons également étudié la nature et la diversité des différentes ressources sémantiques constituées manuellement. Après avoir constaté des différences significatives de taille entre les ressources anglaises et les ressources existant pour le traitement du français, nous avons proposé deux nouvelles méthodes de traduction de ressources très largement utilisées que sont WordNet et FrameNet. Nous essayons actuellement de faire passer les ressources ainsi produites sous licence libre.

**WordNet de noms pour le français** La traduction de WordNet, effectuée uniquement sur les noms, a consisté à utiliser un dictionnaire bilingue pour extraire dans un premier temps toutes les traductions possibles des termes nominaux appartenant à chaque *synset*. Nous avons ensuite proposé des heuristiques pour choisir la meilleure traduction possible de chaque terme source. Notre approche est complètement novatrice dans son utilisation de la structure même de WordNet, c'est-à-dire des relations sémantiques reliant les *synsets* entre eux.

Le système est itératif et commence par traduire les termes considérés monosémiques. Le principe consiste ensuite à laisser le système "deviner" quelle est la meilleure traduction, en analysant les traductions déjà effectuées des *synsets* en relation sémantique. Nous supposons que ces liens fournissent suffisamment d'information pour que l'exercice puisse être résolu par un humain sans trop de difficultés. Aussi, pour permettre au système de résoudre un tel exercice, nous définissons des heuristiques de caractérisation distributionnelle de certaines des relations sémantiques présentes dans WordNet. Le système peut alors exploiter les espaces sémantiques dont nous disposons pour choisir le meilleur candidat de traduction.

Nous avons procédé à une évaluation partiellement automatique en comparant la ressource que nous avons produite avec la ressource Wolf, un WordNet français acquis automatiquement par [Sagot & Fišer 2008]. Wolf n'étant par nature pas exhaustive, nous avons également procédé à l'analyse manuelle de plusieurs échantillons de nos données absentes de Wolf. En considérant uniquement les termes polysémiques, la ressource que nous avons produite contient ainsi plus du double du nombre de paires d'association (terme nominal, *synset*) avec une perte de précision estimée à 10 points, donnant ainsi une précision de l'ordre de 67% au lieu de 77%.

**FrameNet pour le français** La méthode de traduction de FrameNet utilise quant à elle l'hypothèse fondatrice suivante : si une unité lexicale française est la traduction commune de plusieurs unités lexicales source appartenant à la même frame, alors on a une forte confiance en sa présence dans cette frame donnée. Nous proposons donc un score de confiance fondé sur cette hypothèse ainsi

que plusieurs variations visant à pallier différents biais introduits par la structure non homogène des frames (notamment par le nombre d'éléments différent selon les frames).

Le système commence par extraire toutes les traductions possibles à partir de dictionnaires bilingues, puis procède à un filtrage en fonction des scores attribuées aux traductions.

Afin de procéder à l'optimisation de certains paramètres ainsi qu'à l'évaluation partielle de notre ressource, nous avons construit deux échantillons de référence à partir des ressources produites non filtrées et de l'union de la ressource produite par [Padó & Pitel 2007]. Les résultats que nous obtenons sont très intéressants : nous obtenons d'une part une ressource dont la précision est estimée à 95% mais de taille plus petite que le FrameNet original de Berkeley avec une proportion de 58%, d'autre part une seconde ressource environ trois fois plus grande que le FrameNet original avec une précision estimée à 70%.

Ces résultats donnent également un meilleur ensemble d'unités lexicales que la ressource produite par [Padó & Pitel 2007], puisqu'à taille équivalente, nous obtenons une précision supérieure de 18 points. En revanche, à la différence de [Padó & Pitel 2007], notre ressource ne fournit aucune information sur la réalisation des rôles.

Enfin, afin d'enrichir la ressource française obtenue, nous avons également appliqué l'algorithme de classification multi-représenté de [Kriegel *et al.* 2005] à l'ensemble des noms du vocabulaire français figurant dans nos espaces sémantiques. Cet enrichissement a ainsi fourni environ 9000 unités lexicales nominales supplémentaires à la ressource de précision 95% obtenue précédemment (pour une précision finale estimée à 86%).

### 7.1.2 Algorithmes

Au-delà de la constitution de ressources, nos travaux ont également porté sur les tâches d'analyse sémantique elles-mêmes, consistant d'une part en la désambiguïsation lexicale, et d'autre part, en l'annotation en rôles sémantiques. Outre les espaces sémantiques précédemment exploités, les modules développés pour la résolution de ces deux tâches nécessitent l'utilisation d'un analyseur syntaxique de dépendances. Ici, l'analyseur utilisé est toujours LIMA ([Besançon & Chalendar (de) 2005]). Il est recommandé d'utiliser le même analyseur syntaxique que celui ayant servi à la constitution des espaces sémantiques. Dans le cas contraire, une mise en correspondance des différentes relations sera nécessaire.

#### 7.1.2.1 Désambiguïsation lexicale

Nous avons posé le problème de la désambiguïsation lexicale comme un problème de classification supervisé (multi-représenté) exploitant les représentations des mots appartenant aux clusters comme

données d'apprentissage. Ainsi, en exploitant une méthode supervisée sur des données acquises sans supervision, le système lui-même est indépendant de tout corpus d'apprentissage annoté en sens.

D'après l'évaluation menée sur les données de la campagne Romanseval [Segond 2000], notre système n'apparaît que peu performant. Cependant, nous avons montré que la plupart des mots comptés comme étant mal désambiguïsés sont en réalité correctement annotés par nos sens issus des clusters. C'est en revanche la mise en correspondance automatique avec les sens de référence de la campagne qui a fait défaut.

Pour vérifier de façon plus quantitative une telle affirmation, nous avons également calculé la V-mesure associée à ces résultats. Cette mesure a été utilisée dans la tâche de désambiguïsation spécifique aux sens de mots induits de la dernière campagne d'évaluation SemEval ([Manandhar *et al.* 2010]).

Bien que les données ne soient pas identiques (que ce soit pour la langue et par conséquent pour le corpus et le référentiel de sens utilisés), nous avons comparé la valeur de cette mesure calculée sur nos résultats à celle des meilleurs systèmes de Semeval, montrant ainsi une meilleure V-mesure pour notre système. Sans prouver que notre système est plus pertinent que les systèmes anglais de Semeval, cela montre en revanche qu'il permet une désambiguïsation pertinente lorsqu'il n'a pas besoin d'être mis en correspondance avec un référentiel de sens manuel.

#### 7.1.2.2 Annotation en rôles sémantiques

La deuxième tâche d'analyse sémantique sur laquelle nous avons travaillé est l'annotation automatique en rôles sémantiques. Hormis le système de [Padó & Pitel 2007] consistant à projeter automatiquement les annotations de la partie anglaise vers la partie française d'une paire de textes parallèles, nous ne connaissons aucun autre système d'annotation en rôles sémantiques pour le français.

En effet, la plupart des systèmes d'annotation sont des systèmes supervisés, et on ne dispose que depuis récemment du corpus produit par [Padó & Pitel 2007]. De plus, celui-ci ne contient que 1000 phrases, ce qui correspond à un peu moins de la moitié du nombre de phrases annotées de FrameNet. Or le corpus de FrameNet est fréquemment critiqué pour l'insuffisance des données qu'il fournit aux systèmes supervisés pour résoudre une tâche aussi complexe. Notre méthode est quant à elle semi-supervisée dans le sens où elle exploite les traductions des têtes de syntagme issues d'un corpus annoté.

Le cœur de notre méthode réside dans l'introduction d'informations issues des espaces distributionnels syntaxiques. À notre connaissance, seuls [Swier & Stevenson 2004] et [Abend *et al.* 2009] ont utilisé une approche en partie semblable à la nôtre mais ceci dans un cadre plus restreint comme nous l'avons montré à la section 6.1.3.

Nous avons développé l'idée directrice de notre système sur des exemples simples. L'évaluation automatique laisse supposer que cette méthode est trop naïve pour des cas réels. L'introspection manuelle des résultats montre néanmoins que le corpus utilisé comme référence dans l'évaluation ne couvre qu'une partie incomplète de l'ensemble des possibilités d'annotations, biaisant très fortement l'évaluation.

#### 7.1.2.3 Application à la Recherche d'Information

Ayant pour objectif d'intégrer ces modules d'analyse sémantique dans des systèmes appliqués de recherche d'information afin d'évaluer leur apport, nous avons finalement spécifié leur mode d'intégration suivant deux axes présentés aux sections 4.3 et 6.3.

Que cela concerne la désambiguïsation ou l'annotation en rôles, nous considérons l'utilisation de l'information sémantique issue de l'analyse dès la phase d'indexation. Cela permet de pouvoir offrir aux utilisateurs des fonctionnalités de recherche plus riches, comme le contrôle du sens ou l'utilisation de prédicats. Nous recommandons cependant de laisser à l'utilisateur un certain contrôle sur l'interprétation automatique de sa requête par l'interaction avec une interface spécifique.

Nous pensons également que l'utilisation de tels modules permettrait d'améliorer les systèmes de recherche de réponses à des questions posées en langue naturelle. L'utilisation des rôles sémantiques nous a semblé particulièrement adaptée comme nous l'avons expliqué à la section 6.3.

Enfin, en ce qui concerne l'extension de requête, nous pensons que trop d'étapes séparent encore nos différentes ressources et modules de la génération de termes d'extension pertinents.

Ces spécifications concernant l'indexation sémantique et l'extraction de réponses ne sont qu'à l'état d'ébauche. La mise en place de tels prototypes permettrait d'étudier plus formellement leur intérêt.

## 7.2 Discussions et perspectives

Les systèmes automatiques pouvant sans cesse être améliorés, nous nous penchons maintenant sur les thèmes qui nous ont semblé les plus importants à poursuivre et indiquons différentes pistes à ce sujet.

### 7.2.1 Ressources

Dans un premier temps, nous pensons que la méthode de traduction proposée pour la ressource WordNet pourrait être améliorée à la fois par une exploitation conjointe des différents relations (au lieu de n'exploiter qu'un type de relation par itération), et d'autre part par la fixation d'un seuil de non validation pour les raisons déjà invoquées à la section 3.1.2.5. En outre, la méthode ayant

produit des résultats intéressants pour les noms, nous envisageons maintenant de reproduire une expérience similaire sur les verbes et les adjectifs.

Par ailleurs, les clusters de sens n'ayant été produits pour l'instant que sur les mots évalués lors de la campagne Romanseval, nous pouvons à présent produire ces clusters de sens pour l'ensemble du vocabulaire afin de pouvoir les utiliser dans notre algorithme de désambiguïsation.

Enfin, afin d'éclaircir les difficultés rencontrées avec l'évaluation Romanseval lors de la mise en correspondance des sens acquis automatiquement avec les sens manuels, il serait intéressant de tenter de mettre en correspondance nos sens induits à la section 3.2 avec les *synsets* de la ressource de type WordNet que nous avons créée à la section 3.1. En effet, celle-ci contient une information sémantique structurée. Nous pourrions ainsi tenter de classifier nos clusters de sens parmi les *synsets* en utilisant les différentes heuristiques de caractérisation des relations sémantiques définies lors de la traduction de WordNet.

### 7.2.2 Algorithmes

La perspective la plus immédiate pour l'analyse en rôles sémantiques est de procéder à une analyse plus poussée de nos résultats afin d'obtenir à la fois une estimation plus précise de la précision du système et une connaissance plus complète des difficultés rencontrées par notre méthode.

D'autre part, certains choix dans cette première étude ont été faits de façon arbitraire et nécessiteraient d'être établis de façon plus formelle. Cela concerne notamment la façon de combiner les scores ainsi que les seuils de confiance.

Enfin, nous avons établi en 6.1.3 le parallèle entre notre méthode et le modèle plus formel de [Swier & Stevenson 2004]. La perspective la plus prometteuse que nous voyons aujourd'hui pour l'usage de l'information issue des espaces syntactico-sémantiques, est d'envisager une intégration dans un modèle plus proche des modèles d'apprentissage supervisés que le système que nous avons proposé.

Par ailleurs, la combinatoire générée par notre approche théorique pose notamment des problèmes de performance qui sont incompatibles avec une utilisation en temps réel d'une telle analyse. Il conviendrait de fixer formellement des seuils permettant de limiter les temps de traitements et/ou de proposer une implémentation optimisée de cet algorithme.

Concernant ce dernier point, nous tenons à soulever le fait qu'il est regrettable que les campagnes d'évaluation en analyse sémantique ne proposent presque jamais de critère de classement sur la performance. Bien que nous comprenions qu'une telle évaluation soit difficile à mettre en place d'une part, et que d'autre part la recherche d'une solution au plus proche de la vérité soit primordiale, nous soulignons le fait que les temps de traitement de tâches aussi appliquées revêtent également

une importance capitale. Au vu de la littérature étudiée, nous n'avons que très peu d'information concernant l'état de l'art à ce sujet.

Enfin, nous aimerions aussi mettre en place une expérience consistant à évaluer l'influence de la qualité des analyses syntaxiques en entrée de nos systèmes. Cela consisterait à reproduire les expériences menées à la fois sur la tâche de désambiguïsation lexicale et la tâche d'annotation en rôles sémantiques, en utilisant cette fois un autre analyseur syntaxique. Cela permettrait de comparer les résultats ainsi produits avec les résultats obtenus avec l'analyseur syntaxique que nous avions utilisé jusqu'à présent.

Cela ne nécessite pas nécessairement une reconstruction des espaces sémantiques. Pour s'abstraire d'une telle tâche, il conviendrait d'établir une correspondance manuelle entre les relations syntaxiques des espaces sémantiques et les relations syntaxiques produites par l'analyseur testé.

### 7.2.3 Applications

Enfin, l'intégration des algorithmes d'analyse dans des systèmes d'information n'a pas encore été effectuée mais nous avons décrits dans ce manuscrit aux sections 4.3 et 6.3 les détails des prototypes que nous souhaitions implémenter.

### 7.2.4 L'avenir des espaces distributionnels

Nous avons au cours de ces travaux eu l'occasion de découvrir et imaginer d'autres facettes des espaces sémantiques ainsi que d'autres types d'espaces distributionnels. Nous pensons que l'exploration des espaces distributionnels dont l'origine remonte à l'hypothèse soutenue par [Harris 1985] est encore loin d'être achevée. Nous mentionnons ici les pistes d'exploration dont le potentiel nous a paru le plus grand dans le cadre de la représentation du sens et de l'accès à l'information.

#### 7.2.4.1 Compositionalité dans les espaces sémantiques

En sémantique formelle, le principe de compositionalité est défini par [Partee 1984] de la façon suivante :

> L'interprétation d'une expression complexe est une fonction de l'interprétation de ses parties et de la manière dont elles sont assemblées.

Pour donner un exemple, ce qui différencie l'association prédicative *prendre sa retraite* de l'association *prendre son bain* réside dans le sens du deuxième mot. C'est cette composition qui donne alors son sens au prédicat *prendre* qui hors contexte ne signifie que peu de choses.

Le défi que tentent de relever les travaux de [Landauer & Dumais 1997], [Kintsch 2001], [Widdows 2008], [Erk & Padó 2008], [Giesbrecht 2009] ou [Mitchell & Lapata 2010] consiste à capturer la

position vectorielle de ce type d'expressions multimots dans les espaces sémantiques. Ils définissent pour cela des opérateurs vectoriels entre deux mots d'une composition (somme, produit, ou opérateur plus complexe). Le résultat de l'opérateur sur ces deux mots représente alors l'élément composé. Les recherches dans cette direction sont encore récentes et sont en plein développement.

En imaginant que l'on applique un de ces opérateurs sur les mots de nos espaces sémantiques, nous pourrions intégrer les expressions multimots ainsi produites à la fois dans la construction de nos sens induits, dans le choix des traductions de WordNet (à la place d'utiliser les têtes de syntagmes comme nous le faisions jusqu'ici), ainsi que pour l'enrichissement des unités lexicales de FrameNet.

#### 7.2.4.2 Espaces désambiguïsés

Un espace sémantique est construit sur des mots. Certains mots étant ambigus, ceux-ci se retrouvent situés dans l'espace, à un endroit indéterminé entre les différents sens qu'ils représentent. Une idée proposée par Guillaume Pitel est de construire des espaces sémantiques sur un corpus lexicalement désambiguïsé. Dans ce cadre-là, on ne considère plus comme éléments de l'espace les mots eux-mêmes, mais le sens de ceux-ci. Dès lors que les éléments et les contextes d'un tel espace sont désambiguïsés, la position des éléments dans l'espace devient beaucoup plus précise.

La constitution d'un espace désambiguïsé suppose de connaître à l'avance les différents sens des mots et de disposer d'un annotateur automatique le plus précis possible. On avait supposé lors de la constitution des espaces syntactico-sémantiques que les mauvaises analyses syntaxiques seraient négligeables au vu de la quantité de données traitées. De la même façon, on peut supposer que si les mauvaises désambiguïsations lexicales ne sont pas trop nombreuses, la quantité de données acquises donnera des matrices cohérentes malgré le bruit produit par de mauvaises désambiguïsations.

La dernière difficulté que nous voyons dans la construction d'un tel espace provient de la rareté de certains sens en corpus. Il faudrait donc utiliser un corpus d'une taille plus grande ou surtout plus varié que pour la construction d'un espace sémantique classique. En effet, on peut déduire de l'hypothèse *One sense per discourse* de [Yarowsky 1993] qu'un corpus plus varié donnera lieu à des occurrences de sens plus variées pour chaque mot.

#### 7.2.4.3 Espaces de paraphrases

Dans le but de constituer des patrons d'acquisition de relations, [Bhagat & Ravichandran 2008] construisent un espace distributionnel de paraphrases (à très grande échelle). En effet, ils considèrent que les éléments à projeter dans l'espace sont les chemins syntaxiques (incluant les lexèmes) reliant chaque mot à un autre dans une phrase. Les deux mots en question sont considérés comme contexte gauche et droit du chemin ainsi constitué. Dans ce cadre, les auteurs définissent alors une représentation vectorielle de chaque chemin. Celle-ci est définie comme étant la concaténation du vecteur représentant les nombres d'occurrences de ses contextes gauches et du vecteur représentant

les nombres d'occurrences de ses contextes droits. Les chemins inverses sont également inclus dans l'espace avec une inversion des contextes gauches et droits.

Disposant ainsi d'une représentation vectorielle de tous les chemins syntaxiques rencontrés, ils sont alors capables de rechercher les plus proches voisins de chaque chemin syntaxique afin de découvrir ses paraphrases.

Les paraphrases peuvent s'avérer très utiles en recherche d'information, notamment pour proposer des reformulations de question.

#### 7.2.4.4 Espaces multilingues

Un certain nombre de tâches crosslingues peuvent également trouver un intérêt dans les espaces sémantiques multilingues. Dans ce paradigme, des mots de langues différentes se trouvent projetés dans un même espace et disposent ainsi d'une similarité sémantique. Ceci est notamment intéressant dans le cadre de la recherche crosslingue.

La constitution d'espaces multilingues peut se faire à partir de dictionnaires bilingues. Elle est alors faite par traduction brute des contextes sans tenir compte de leur polysémie potentielle et en espérant que la quantité de données analysées suffira à combler ce biais.

[Peirsman & Padó 2010] proposent également une méthode de constitution de tels espaces par un *bootstrapping* exploitant les cognats comme premiers éléments de la constitution. Ils exploitent ensuite cet espace bimodal pour apprendre des préférences de sélection sémantique (*selectional preferences*).

#### 7.2.4.5 Espaces multimodaux

Nous appelons espace multimodal un espace regroupant des informations contextuelles de différentes modalités. Il peut s'agir de texte, mais aussi d'images, de vidéos, de son... Un tel espace se construit à partir d'un corpus lui-même multimodal, c'est-à-dire comprenant différents médias associés les uns les autres pour former un seul et même document. Il peut par exemple s'agir d'un corpus dans lequel pour chaque image on trouve un paragraphe de texte décrivant l'image.

En effet, en traitement du signal, il existe des descripteurs locaux correspondant au pendant du vocabulaire dans le traitement du texte. Par exemple, dans le domaine du traitement d'image, la similarité entre deux images peut être calculée à partir de ce qu'on appelle des descripteurs locaux ou *mots visuels*. Ces descripteurs sont construits à partir de vecteurs calculés sur des points saillants de l'image et rapprochés ensuite d'un vocabulaire visuel prédéfini. Chaque image est constituée d'un certain nombre de descripteurs locaux, de la même façon qu'un texte est constitué d'un certain nombre de mots. On parle ainsi de vocabulaire visuel.

À partir d'un corpus multimodal, nous pouvons donc construire l'espace distributionnel formé par les cooccurrences des descripteurs locaux de chacun des médias du document (mots textuels et mots visuels dans le cas d'un corpus texte-image).

Nous avons initié et dirigé un travail de stage pour construire un tel espace et tenter de vérifier l'hypothèse sous-jacente concernant la possibilité d'aller vers une recherche de documents cross-modale ([Walkowiak 2010]). En effet, la recherche des plus proches voisins d'un mot textuel pouvant mener à la découverte de mot visuels et vice-versa, nous pouvons espérer qu'en cherchant les plus proches voisins de tous les mots d'une requête textuelle, un système pourrait parvenir à extraire une image correspondante. L'inverse est également imaginable.

Si un tel modèle de représentation de données est possible, alors on assistera à une réduction conséquente du fossé sémantique séparant les médias non textuels des concepts symboliques véhiculés par le langage.

En pratique, cela s'avère difficile à mettre en oeuvre. Nous voyons deux problématiques à résoudre dans un premier temps. La première réside dans le fait qu'il faut parvenir à déterminer une taille optimale de vocabulaire visuel utilisé, mais on ne sait pas définir ce qui est optimal. D'autre part il faut également déterminer un moyen de pondérer les occurrences des mots visuels, qui sont d'une fréquence beaucoup plus élevée que les mots textuels. En effet, la fréquence d'un mot visuel donné peut s'élever à plusieurs centaines d'occurrences dans une même image, ce qui est exceptionnel dans le cas d'un mot textuel. Les travaux de stage n'ont pas été menés suffisamment loin pour pouvoir conclure sur la possibilité de recherche cross-modale.

L'idée d'aller vers l'exploration d'un tel type d'espaces est confortée par le fait que d'autres personnes s'y intéressent. En particulier, [Feng & Lapata 2010] construisent un espace de ce type et montrent que les similarités rapportées entre les mots textuels sont plus précises en prenant en compte également les traits visuels que sans les prendre en compte. Leur première étude ne va pas plus loin dans l'exploitation de la bimodalité de l'espace.

### 7.2.5 Synthèse

Pour conclure sur l'ensemble des travaux présentés, nous souhaitons mettre en avant le fait que chacune des tâches effectuées exploite en permanence les espaces distributionnels syntaxiques. Ceci est vrai tant dans la construction des ressources que dans les algorithmes fonctionnels eux-mêmes. La pertinence des résultats obtenus à chaque étape montre que cette modélisation joue un rôle fondamental à la fois pour traduire et enrichir des ressources manuelles mais également pour combler le manque de corpus annoté (par sens ou par rôles) et parvenir à l'analyse sémantique de textes français.

Plus généralement, nous pensons que l'exploration des espaces distributionnels et la création de nouveaux modèles (en particulier multimédia/multimodal) est une source fascinante d'information, notamment pour la recherche de documents.

D'après Ray Kurzweil [1], en 2029 les machines interprèteront le texte de façon suffisamment proche de ce qu'un humain ferait, pour être capable de passer le test de Turing ([Kurzweil 2005]). L'analyse du langage et plus particulièrement l'analyse sémantique automatique est une brique considérable nécessaire à une telle réalisation.

Nous avons dans ce travail proposé et produit un nombre conséquent de ressources et outils pour l'analyse automatique du sens de textes en langue française. Nous espérons avoir ainsi contribué à notre échelle aux efforts ayant pour objectif commun de rendre le texte interprétable par les machines.

1. un des futuristes désigné par l'Agence Nationale d'Ingénierie Américaine pour identifier les grands défis du 21$eme$ siècle

# Bibliographie

[Abend & Rappoport 2010] Omri Abend et Ari Rappoport. Fully Unsupervised Core-Adjunct Argument Classification. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010.

[Abend *et al.* 2009] Omri Abend, Roi Reichart et Ari Rappoport. Unsupervised argument identification for Semantic Role Labeling. In ACL-IJCNLP'09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, volume 1, pages 28–36, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[Achtert *et al.* 2006] Elke Achtert, Hans-Peter Kriegel, Alexey Pryakhin et Matthias Schubert. Clustering Multi-Represented Objects Using Combination Trees. In Proceedings 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2006), Singapore, 2006.

[Agichtein & Gravano 2000] Eugene Agichtein et Luis Gravano. Snowball: extracting relations from large plain-text collections. In DL '00: Proceedings of the fifth ACM conference on Digital libraries, pages 85–94, New York, NY, USA, 2000. ACM.

[Agirre & Edmonds 2007] Eneko Agirre et Philip Edmonds, editeurs. Word sense disambiguation - algorithms and applications. Springer, 2007.

[Agirre & Rigau 1996] Eneko Agirre et German Rigau. Word Sense Disambiguation using Conceptual Density. In Proceedings of COLING-96, 1996.

[Agirre & Soroa 2007] Eneko Agirre et Aitor Soroa. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In Proceedings of the Fourth International Workshop on Semantic Evaluations, pages 7–12, Prague, Czech Republic, June 2007. ACL.

[Agirre *et al.* 2009] Eneko Agirre, Giorgio Maria Di Nunzio, Thomas Mandl et Arantxa Otegi. CLEF 2009 Ad Hoc Track Overview: Robust-WSD Task. In Proceedings of CLEF 2009, 2009.

[Ahlswede & Lorand 1993] T. Ahlswede et D Lorand. Word sense disambiguation by human subjects: Computational and psycholinguistic applications. In Proceedings of the Workshop on Acquisitions of Lexical Knowledge from Text, pages 1–9, Columbus, Ohio, 1993.

[Alilaghatta 2006] Vijay Alilaghatta. DBpedia and (Open-)Cyc (http://wiki.dbpedia.org/OpenCyc?v=uy6), 2006.

[Amaral *et al.* 2004] Carlos Amaral, Dominique Laurent, André Martins, Afonso Mendes et Cláudia Pinto. Design and Implementation of a Semantic Search Engine for Portuguese. In Proceedings of LREC 2004, 2004.

[Anick *et al.* 2008] Peter Anick, Vijay Murthi et Shaji Sebastian. Similar Term Discovery using Web Search. In Proceedings of LREC 2008, Marrakech, 2008.

[Ankerst *et al.* 1999] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel et Jörg Sander. OPTICS: Ordering Points to Identify the Clustering Structure. In ACM Press, editeur, Proceedings ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'99 ), pages 49–60, New York, May 1999.

[Artiles *et al.* 2009] Javier Artiles, Enrique Amigó et Julio Gonzalo. The role of named entities in web people search. In EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 534–542, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[Atkins 1994] Sue Atkins. I Don't Believe in Word Senses. Past President, European Association for Lexicography; General Editor, Collins-Robert English/French Dictionary; Lexicographical Adviser, Oxford University Press - responding to a discussion which assumed discrete and disjoint word senses. Cited by [Kilgarriff 97], Octobre 1994.

[Attar & Fraenkel 1977] R. Attar et A. S. Fraenkel. Local Feedback in Full-Text Retrieval Systems. J. ACM, vol. 24, no. 3, pages 397–417, 1977.

[Auer *et al.* 2008] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak et Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In Proceedings of 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007), pages 722–735, November 2008.

[Auger & Barrière 2008] Alain Auger et Caroline Barrière. Pattern-based approaches to semantic relation extraction: A state-of-the-art. Terminology, vol. 14, no. 1, pages 1–19, 2008.

[Baker *et al.* 1998] Collin F. Baker, Charles J. Fillmore et John B. Lowe. The Berkeley FrameNet Project. In Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics, pages 86–90, Montréal, Canada, 1998.

[Baker *et al.* 2007] Collin F. Baker, Michael Ellsworth et Katrin Erk. SemEval-2007 task 19: Frame Semantic Structure Extraction. In SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations, pages 99–104, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

[Bansal *et al.* 2004] Nikhil Bansal, Avrim Blum et Shuchi Chawla. Correlation Clustering. Machine Learning, vol. 56, no. 1-3, pages 89–113, 2004.

[Barbu & Barbu Mititelu 2005] Eduard Barbu et Verginica Barbu Mititelu. Automatic Building of Wordnets. In Proceedings of RANLP 2005, 2005.

[Basili *et al.* 2009] Roberto Basili, Diego Cao, Danilo Croce, Bonaventura Coppola et Alessandro Moschitti. Cross-Language Frame Semantics Transfer in Bilingual Corpora. In Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), pages 332 – 345, Mexico City, Mexico, 2009.

[Bejan & Hathaway 2007] Cosmin Adrian Bejan et Chris Hathaway. UTD-SRL: A Pipeline Architecture for Extracting Frame Semantic Structures. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 460–463,, Prague, June 2007.

[Bejan *et al.* 2004] Cosmin Adrian Bejan, Alessandro Moschitti, Paul Morărescu, Gabriel Nicolae et Sanda Harabagiu. Semantic Parsing Based on FrameNet. In Proceedings of SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, july 2004.

[Bejoint & Thoiron 1996] Henri Bejoint et Philippe Thoiron. Les dictionnaires bilingues. Duculot - De Boeck, 1996.

[Bentivogli *et al.* 2009] Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo et Bernardo Magnini. The Fifth PASCAL Recognizing Textual Entailment Challenge. In TAC 2009 Workshop, Gaithersburg, Maryland, USA, 2009.

[Besançon & Chalendar (de) 2005] Romaric Besançon et Gaël Chalendar (de). L'analyseur syntaxique de LIMA dans la campagne d'évaluation Easy. In Actes de TALN 05, 2005.

[Bezdek 1981] James C. Bezdek. Pattern recognition with fuzzy objective function algorithms. Kluwer Academic Publishers, Norwell, MA, USA, 1981.

[Bhagat & Ravichandran 2008] Rahul Bhagat et Deepak Ravichandran. Large Scale Acquisition of Paraphrases for Learning Surface Patterns. In Proceedings of ACL-08: HLT, page 674–682, Columbus, Ohio, USA, June 2008.

[Bhogal *et al.* 2007] J. Bhogal, A. Macfarlane et P. Smith. A review of ontology based query expansion. Information Processing and Management, vol. 43, no. 4, pages 866–886, 2007.

[Bickel & Scheffer 2004] Steffen Bickel et Tobias Scheffer. Multi-View Clustering. In Proceedings of the Fourth IEEE International Conference on Data Mining, pages 19–26. IEEE Computer Science, 2004.

[Bordag 2006] Stefan Bordag. Word Sense Induction: Triplet-Based Clustering and Automatic Evaluation. In Proceedings of the 11th EACL, pages 137–144, Trento, Italy, 2006.

[Brin & Page 1998] Sergey Brin et Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, vol. 30, no. 1-7, pages 107 – 117, 1998. Proceedings of the Seventh International World Wide Web Conference.

[Briscoe & Carroll 1997] E. Briscoe et J. Carroll. Automatic extraction of subcategorization from corpora. In Proceedings of the 5th ACL Conference on Applied Natural Language Processing, page 356–363, Washington, DC, 1997.

[Brown *et al.* 1991] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra et Robert L. Mercer. Word-sense disambiguation using statistical methods. In Proceedings of ACL, pages 264–270, Berkeley, California, USA, 1991.

[Buckley *et al.* 1995] Chris Buckley, Gerard Salton, James Allan et Amit Singhal. Automatic Query Expansion Using SMART : TREC 3. In In Proceedings of The third Text REtrieval Conference (TREC-3, pages 69–80, 1995.

[Burchardt *et al.* 2006] Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó et Manfred Pinka. The SALSA corpus: A German corpus resource for lexical semantics. In Proceedings of LREC 2006, pages 969–974, Genoa, Italy, 2006.

[Chalendar (de) *et al.* 2002] Gaël Chalendar (de), Tiphaine Dalmas, Faïza Elkateb-Gara, Olivier Ferret, Brigitte Grau, Martine Hurault-Plantet, Gabriel Illouz, Laura Monceaux, Isabelle Robba et Anne Vilnat. The question answering system QALC at LIMSI: experiments in using Web and WordNet. In In Proceedings of the Eleventh Text REtrieval Conference (TREC, pages 407–416, 2002.

[Chalendar (de) 2001] Gaël Chalendar (de). SVETLAN, un système de structuration du lexique guidé par la détermination automatique du contexte thématique. PhD thesis, Université Paris XI, Orsay, 2001.

[Chambers 1728] Ephraïm Chambers, editeur. Cyclopaedia ou dictionnaire universel des arts et des sciences. Chambers, Ephraïm, 1728.

[Chan *et al.* 2007] Yee Seng Chan, Hwee Tou Ng et Zhi Zhong. NUS-PT: Exploiting Parallel Texts for Word Sense Disambiguation in the English All-Words Tasks. In Proceedings of the 4th International Workshop on Semantic Evaluations - SemEval'07, 2007.

[Charikar 2002] Moses Charikar. Similarity Estimation Techniques from Rounding Algorithms. In Proceedings of the 34th Annual ACM Symposium on Theory of Computing, 2002.

[Chibout 1998] Karim Chibout. La polysémie lexicale : observations linguistiques, modélisation informatique, études ergonomique et psycholinguistique. PhD thesis, Université Paris 11, Orsay, 1998.

[Cimiano & Völker 2005] P. Cimiano et J. Völker. Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery. In Proceedings of the 10th International Conference

on Applications of Natural Language to Information Systems (NLDB), Lecture Notes in Computer Science, Alicante, Spain, June 2005.

[Cover & Hart 1967] T. M. Cover et P. E. Hart. Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, vol. IT-13, no. 1, pages 21–27, january 1967.

[Crabtree *et al.* 2007] Daniel Crabtree, Peter Andreae et Gao Xiaoying. Exploiting Underrepresented Query Aspects for Automatic Query Expansion. In The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007), San Jose, California, United States, August 2007. ACM SIGKDD.

[Crestan *et al.* 2003] Éric Crestan, Marc El-Bèze et Claude Loupy (de). Peut-on trouver la taille de contexte optimale en désambiguïsation sémantique ? In Actes de TALN 2003, pages 85–94, Batz-sur-Mer, juin 2003.

[Cruse 1995] Alan Cruse. Polysemy and Related Phenomena from a Cognitive Linguistic Viewpoint. Computational Lexical Semantics, pages 33–49, 1995.

[Curran 2004] James R. Curran. From Distributional to Semantic Similarity. PhD thesis, University of Edinburgh, 2004.

[Dagan & Itai 1994] Ido Dagan et Alon Itai. Word Sense Disambiguation Using a Second Language Monolingual Corpus. Computational Linguistics, pages 563–596, 1994.

[Dagan *et al.* 1991] Ido Dagan, Alon Itai et Ulrike Schwall. Two languages are more informative than one. In Proceedings of ACL 29, pages 130–137, 1991.

[Dang & Palmer 2005] Hoa Trang Dang et Martha Palmer. The Role of Semantic Roles in Disambiguating Verb Senses. In Proceedings of the 43rd Meeting of the Association for Computational Linguistics (ACL-05), Ann Arbor, pages 26–28, 2005.

[Dang *et al.* 2007] Hoa Trang Dang, Diane Kelly et Jimmy J. Lin. Overview of the TREC 2007 Question Answering Track. In Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, November 2007.

[Das *et al.* 2010] Dipanjan Das, Nathan Schneider, Desai Chen et Noah A. Smith. Probabilistic Frame-Semantic Parsing. In Proceedings of NAACL 2010, Los Angeles, USA, 2010.

[Davidson 1967] D. Davidson. The logic of decision and action, chapitre The logical Form of Action Sentences, pages 81–120. University of Pittsburgh Press, 1967.

[Dempster *et al.* 1977] A. P. Dempster, N. M. Laird et D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B, vol. 39, no. 1, pages 1–38, 1977.

[Deschacht & Moens 2009] Koen Deschacht et Marie-Francine Moens. Semi-supervised semantic role labeling using the latent words language model. In EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 21–29, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[Diderot & D'Alembert 1751] Denis Diderot et Jean le Rond D'Alembert. Encyclopédie ou dictionnaire raisonné des sciences, des arts et des métiers. Le Breton, André, 1751.

[Dini *et al.* 2000] Luca Dini, Vittorio di Tomaso et Frédérique Segond. GINGER II: An Example-Driven Word Sense Disambiguator. Computer and the Humanities, vol. 34, pages 121–126, 2000.

[Dorow & Widdows 2003] Beate Dorow et Dominic Widdows. Discovering Corpus-Specific Word Senses. In EACL 2003, 2003.

[Dowty 1991] David. R. Dowty. Thematic Proto-Roles and Argument Selection. Language, vol. 67, pages 547–619, 1991.

[Duin 2002] Robert P. W. Duin. The combining classifier: to train or not to train? In Pattern Recognition, 2002. Proceedings. 16th International Conference on, volume 2, pages 765–770 vol.2, 2002.

[Edmonds 2005] Philip Edmonds. The elsevier encyclopedia of language and linguistics, 2nd ed., chapitre Lexical Disambiguation. Oxford: Elsevier, 2005.

[Elshamy *et al.* 2010] Wesam Elshamy, Doina Caragea et William Hsu. KSU KDD: Word Sense Induction by Clustering in Topic Space. In Proceedings of the 5th International Workshop on Semantic Evaluation, pages 367–370, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[Embarek & Ferret 2008] Mehdi Embarek et Olivier Ferret. Learning patterns for building resources about semantic relations in the medical domain. In Proceedings of LREC 08, 2008.

[Erk & Padó 2006] Katrin Erk et Sebastian Padó. Shalmaneser - A toolchain for shallow semantic parsing. In Proceedings of LREC 2006, 2006.

[Erk & Padó 2008] Katrin Erk et Sebastian Padó. A Structured Vector Space Model for Word Meaning in Context. In Proceedings of EMNLP 2008, 2008.

[Ertöz *et al.* 2001] L. Ertöz, M. Steinbach et V. Kumar. Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach. In Proceedings of Workshop on Text Mining, First SIAM International Conference on Data Mining, Chicago, IL, 2001.

[Ester *et al.* 1996] Martin Ester, Hans-Peter Kriegel, Jörg Sander et Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In AAAI Press, editeur, Proceedings 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD'96 ), pages 226–231, Menlo Park, CA, August 1996.

[Etzioni *et al.* 2004] Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld et Alexander Yates. Web-Scale Information Extraction in KnowItAll, 2004.

[Eynde (van den) & Mertens 2003] Karel Eynde (van den) et Piet Mertens. La valence: l'approche pronominale et son application au lexique verbal. Journal of French Language Studies, vol. 13, pages 63–104, 2003.

[Falk & Gardent 2010] Ingrid Falk et Claire Gardent. Bootstraping a Classification of French Verbs Using Formal Concept Analysis. In Proceedings of Interdisciplinary workshop on verbs, Pisa, Italy, 2010.

[Falk *et al.* 2009] Ingrid Falk, Claire Gardent, Evelyne Jacquey et Fabienne Venant. Sens, synonymes et définitions. In Proceedings of TALN 2009, 2009.

[Fellbaum 1998] Christiane Fellbaum, editeur. Wordnet : An electronic lexical database. MIT Press, 1998.

[Feng & Lapata 2010] Yansong Feng et Mirella Lapata. Visual Information in Semantic Representation. In Proceedings of the 2010 Annual Conference of the North American Chapter of the ACL, page 91–99, 2010.

[Fernandez *et al.* 2002] Ana Fernandez, Gloria Vazquez, Patrick Saint-Dizier, Farah Benamara et Mouna Kamel. The VOLEM project: a framework for the construction of advanced multilingual lexicons. In Language Engineering Conference, 2002. Proceedings, pages 89–98, 2002.

[Ferret 1998] Olivier Ferret. ANTHAPSI : un système d'analyse thématique et d'apprentissage de connaissances pragmatiques fondé sur l'amorçage. PhD thesis, UNIVERSITE DE PARIS-SUD, décembre 1998.

[Ferret 2004] Olivier Ferret. Discovering word senses from a network of lexical cooccurrences. In COLING '04: Proceedings of the 20th international conference on Computational Linguistics, pages 1326–1332, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[Ferret 2010] Olivier Ferret. Similarité sémantique et extraction de synonymes à partir de corpus. In Actes de TALN 2010, Montréal, Canada, July 2010.

[Fillmore 1968] Charles J. Fillmore. Universals in linguistic theory, chapitre The Case for Case, pages 1–88. Holt, Rinehart, and Winston, New York, 1968.

[Fillmore 1976] Charles J. Fillmore. Frame semantics and the nature of language. Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, vol. 280, pages 20–32, 1976.

[Fleischman *et al.* 2003] Michael Fleischman, Namhee Kwon et Eduard Hovy. Maximum entropy models for FrameNet classification. In Proceedings of the 2003 conference on Empirical methods in natural language processing, pages 49–56, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[Francopoulo *et al.* 2006] Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet et Claudia Soria. Lexical Markup Framework (LMF). In International

Conference on Language Resources and Evaluation - LREC 2006, Gênes/Italie, 2006. elra. LIRICS.

[Fred & Jain 2002] Ana L. N. Fred et Anil K. Jain. Data Clustering Using Evidence Accumulation. In ICPR '02: Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 4, page 40276, Washington, DC, USA, 2002. IEEE Computer Society.

[Furnas *et al.* 1987] G. W. Furnas, T. K. Landauer, L. M. Gomez et S. T. Dumais. The vocabulary problem in human-system communication. In Communications of the ACM, volume 30, pages 964–971. ACM, November 1987.

[Fürstenau & Lapata 2009a] Hagen Fürstenau et Mirella Lapata. Graph Alignment for Semi-Supervised Semantic Role Labeling. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMLNP 09), pages 11–20, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[Fürstenau & Lapata 2009b] Hagen Fürstenau et Mirella Lapata. Semi-Supervised Semantic Role Labeling. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 09), pages 220–228. Association for Computational Linguistics, 2009.

[Gale *et al.* 1992a] William A. Gale, Kenneth W. Church et David Yarowsky. A method for disambiguating word senses in a large corpus. Computers and the Humanities, vol. 26, pages 415–439, December 1992.

[Gale *et al.* 1992b] William A. Gale, Kenneth W. Church et David Yarowsky. Work on Statistical Methods for Word Sense Disambiguation. Rapport technique, AT&T Bell Laboratories, 1992.

[Gardent & Lorenzo 2010] Claire Gardent et Alejandra Lorenzo. Identifying sources of weakness in Syntactic Lexicon Extraction. In Proceedings of LREC 2010, Malta, 2010.

[Gaustad & Groningen 2001] Tanja Gaustad et Rijksuniversiteit Groningen. Statistical Corpus-Based Word Sense Disambiguation: Pseudowords vs. Real Ambiguous Words. In In Companion Volume to the Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL/EACL 2001) – Proceedings of the Student Research Workshop, pages 61–66, 2001.

[Giesbrecht 2009] Eugenie Giesbrecht. In Search of Semantic Compositionality in Vector Spaces. In Sebastian Rudolph, Frithjof Dau et Sergei Kuznetsov, editeurs, Conceptual Structures: Leveraging Semantic Technologies, volume 5662 of Lecture Notes in Computer Science, pages 173–184. Springer Berlin / Heidelberg, 2009. 10.1007/978-3-642-03079-6_14.

[Gildea & Jurafsky 2002] Daniel Gildea et Daniel Jurafsky. Automatic labeling of semantic roles. Computational Linguistics, vol. 28, no. 3, pages 245–288, September 2002.

[Gildea & Palmer 2002] Daniel Gildea et Martha Palmer. The necessity of parsing for predicate argument recognition. In ACL '02: Proceedings of the 40th Annual Meeting on Association

for Computational Linguistics, pages 239–246, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[Gildea 2002] Daniel Gildea. Probabilistic Models of Verb-Argument Structure. In Proceedings of the 19th International Conference on Computational Linguistics (COLING-02), 2002.

[Gionis *et al.* 2007] Aristides Gionis, Heikki Mannila et Panayiotis Tsaparas. Clustering aggregation. ACM Trans. Knowl. Discov. Data, vol. 1, no. 1, page 4, 2007.

[Giuglea & Moschitti 2006] Ana-Maria Giuglea et Alessandro Moschitti. Semantic role labeling via FrameNet, VerbNet and PropBank. In ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pages 929–936, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[Goodman 2001] Joshua T. Goodman. A bit of progress in language modeling. Computer Speech & Language, vol. 15, no. 4, pages 403 – 434, 2001.

[Grefenstette 1994] Gregory Grefenstette. Explorations in automatic thesaurus discovery. Kluwer Academic Publishers, Norwell, MA, USA, 1994.

[Grefenstette 2007] Gregory Grefenstette. Conquering language : Using nlp on a massive scale to build high dimensional language models from the web. In Proceedings of the 8th CICLing Conference, pages 35–49, Mexico, Mexico, 2007.

[Grenager & Manning 2006] Trond Grenager et Christopher D. Manning. Unsupervised discovery of a statistical verb lexicon. In EMNLP '06: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pages 1–8, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[Grishman *et al.* 1994] Ralph Grishman, Catherine Macleod et Adam Meyers. Comlex Syntax: Building a Computational Lexicon. In booktitle = Proceedings of COLING-ACL94,, Kyoto, Japon, 1994.

[Gruber 1965] Jeffrey S. Gruber. Studies in Lexical Relations. PhD thesis, MIT, 1965.

[Hajič *et al.* 2009] Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue et Yi Zhang. The CoNLL-2009 shared task: syntactic and semantic dependencies in multiple languages. In CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning, pages 1–18, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[Harnad 1990] Stevan Harnad. The Symbol Grounding Problem. Physica D: Nonlinear Phenomena, vol. 42, pages 335–346, 1990.

[Harris 1985] Zelig Harris. Distributional Structure. In J. J. Katz, editeur, The Philosophy of Linguistics, pages 26–47. Oxford University Press, New York, 1985.

[He & Gildea 2006] Shan He et Daniel Gildea. Self-training and Co-training for Semantic Role Labeling: Primary Report. Rapport technique, The University of Rochester, 2006.

[Hearst 1992] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th conference on Computational linguistics, pages 539–545, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

[Hirst & St Onge 1998] G. Hirst et D. St Onge. Wordnet: An electronic lexical database (language, speech, and communication), chapitre Lexical Chains as representation of context for the detection and correction malapropisms. The MIT Press, May 1998.

[Hirst 1987] Graeme Hirst. Semantic interpretation and the resolution of ambiguity. Cambridge University Press, Cambridge, UK, 1987.

[Hovy *et al.* 2006] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw et Ralph Weischedel. OntoNotes: The 90% solution. In Proceedings of HLT-NAACL 2006, pages 57–60, 2006.

[Indyk & Motwani 1998] Piotr Indyk et R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In Proceedings of 30th STOC, page 604–613, 1998.

[Jackendoff 1972] Ray S. Jackendoff. Semantic interpretation in generative grammar. MIT Press, Cambridge, 1972.

[Jelinek & Mercer 1980] F. Jelinek et R. Mercer. Interpolated estimation of Markov source parameters from sparse data. In Proceedings of the Workshop on Pattern Recognition in Practice, pages 381–397, North-Holland, Amsterdam, 1980.

[Ji *et al.* 2003] Hyungsuk Ji, Sabine Ploux et Eric Wehrli. Lexical Knowledge Representation with Contexonyms, 2003.

[Jing & Croft 1994] Yufeng Jing et W. Bruce Croft. An Association Thesaurus for Information Retrieval. In In RIAO 94 Conference Proceedings, pages 146–160, 1994.

[Johansson & Nugues 2007a] Richard Johansson et Pierre Nugues. LTH: Semantic Structure Extraction using Nonprojective Dependency Trees. In SemEval'07: Proceedings of the 4th International Workshop on Semantic Evaluations, pages 227–230, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

[Johansson & Nugues 2007b] Richard Johansson et Pierre Nugues. Using WordNet to extend FrameNet coverage. In Proceedings of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages, at NODALIDA, pages 27–30, 2007.

[Johnson *et al.* 2003] Christopher Johnson, Miriam Petruck, Collin F. Baker, Michael Ellsworth, Josef Ruppenhofer et Charles J. Fillmore. Framenet: Theory and practice, 2003.

[Jurgens & Stevens 2010] David Jurgens et Keith Stevens. HERMIT: Flexible Clustering for the SemEval-2 WSI Task. In Proceedings of the 5th International Workshop on Semantic Evaluation, pages 359–362, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[Kailing *et al.* 2004] Karin Kailing, Hans-Peter Kriegel, Alexey Pryakhin et Matthias Schubert. Clustering Multi-Represented Objects with Noise. In Proceedings 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2004), Sydney, Australia, 2004.

[Kaisser & Webber 2007] Michael Kaisser et Bonnie Webber. Question answering based on semantic roles. In DeepLP '07: Proceedings of the Workshop on Deep Linguistic Processing, pages 41–48, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

[Kern *et al.* 2010] Roman Kern, Markus Muhr et Michael Granitzer. KCDC: Word Sense Induction by Using Grammatical Dependencies and Sentence Phrase Structure. In Proceedings of the 5th International Workshop on Semantic Evaluation, pages 351–354, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[Kilgarriff & Rosenzweig 2000] Adam Kilgarriff et Joseph Rosenzweig. Framework and results for English SENSEVAL. Computers and the Humanities, vol. 34, no. 1-2, pages 15–48, 2000.

[Kilgarriff 1997] Adam Kilgarriff. I Don't Believe in Word Senses. Computers and the Humanities, 1997.

[Kingsbury & Palmer 2002] Paul Kingsbury et Martha Palmer. From treebank to propbank. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002), pages 1989–1993, 2002.

[Kintsch 2001] Walter Kintsch. Predication. Cognitive Science, vol. 25, page 173–202, 2001.

[Kiryakov & Simov 2000] Atanas K. Kiryakov et Kiril Iv. Simov. Mapping of EuroWordnet Top Ontology into Upper Cyc Ontology, 2000.

[Kittler *et al.* 1998] Josef Kittler, Mohamad Hatef, Robert P. W. Duin et Jiri Matas. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pages 226–239, 1998.

[Kleiber & Tamba 1990] G. Kleiber et I. Tamba. L'hyponymie revisitée : inclusion et hiérarchie. Langages, vol. 25, no. 98, pages 7–32, 1990.

[Koehn 2005] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In Actes de MT Summit X, Phuket, Thaïlande, 2005.

[Kohomban & Lee 2005] Upali S. Kohomban et Wee Sun Lee. Learning semantic classes for word sense disambiguation. In ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 34–41, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[Korhonen 2002] A. Korhonen. Subcategorization Acquisition. PhD thesis, University of Cambridge, 2002.

[Korkontzelos & Manandhar 2010] Ioannis Korkontzelos et Suresh Manandhar. UoY: Graphs of Unambiguous Vertices for Word Sense Induction and Disambiguation. In Proceedings of the

5th International Workshop on Semantic Evaluation, pages 355–358, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[Kotis *et al.* 2006] Konstantinos Kotis, George A. Vouros et Konstantinos Stergiou. Towards automatic merging of domain ontologies: The HCONE-merge approach. Web Semantics: Science, Services and Agents on the World Wide Web, vol. 4, no. 1, pages 60 – 79, 2006.

[Kriegel *et al.* 2005] Hans-Peter Kriegel, Alexey Pryakhin et Matthias Schubert. Multi-represented kNN-Classification for Large Class Sets. In Proceedings of the 10th International Conference on Database Systems for Advanced Applications (DASFAA'05), Beijing, China, 2005.

[Kriegel *et al.* 2008a] Hans-Peter Kriegel, Peter Kunath, Alexey Pryakhin et Matthias Schubert. Distribution-Based Similarity for Multi-Represented Multimedia Objects. In proc. 14th Multimedia Modeling Conference (MMM 2008), Kyoto, Japan, 2008.

[Kriegel *et al.* 2008b] Hans-Peter Kriegel, Peter Kunath, Alexey Pryakhin et Matthias Schubert. MUSE: Multi-Represented Similarity Estimation. In proc. 24th International Conference on Data Engineering (ICDE 2008), Cancún, México, 2008.

[Krovetz & Croft 1992] Robert Krovetz et W. Bruce Croft. Lexical ambiguity and information retrieval. ACM Trans. Inf. Syst., vol. 10, no. 2, pages 115–141, 1992.

[Kuncheva *et al.* 2001] Ludmila I. Kuncheva, James C. Bezdek et Robert P. W. Duin. Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recognition, vol. 34, no. 2, pages 299 – 314, 2001.

[Kurzweil 2005] Ray Kurzweil. The singularity is near. Penguin Books, 2005.

[Landauer & Dumais 1997] T. K. Landauer et S. T. Dumais. Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. Psychological Review, no. 104, 1997.

[Laurent & Séguéla 2005] Dominique Laurent et Patrick Séguéla. QRISTAL, système de Questions-Réponses. In Michèle Jardino, editeur, Actes de TALN 2005 (Traitement automatique des langues naturelles), Dourdan, Juin 2005. ATALA, LIMSI.

[Le *et al.* 2007] Cuong Anh Le, Van-Nam Huynh, Akira Shimazu et Yoshiteru Nakamori. Combining classifiers for word sense disambiguation based on Dempster-Shafer theory and OWA operators. Data & Knowledge Engineering, vol. 63, pages 381–396, 2007.

[Leclère 2005] Christian Leclère. Linguistic informatics - state of the art and the future, chapitre The lexicon-grammar of French verbs: a syntactic database, pages 29–45. Amsterdam/Philadelphia : Benjamins, 2005.

[Lee & Ng 2002] Yoong Keok Lee et Hwee Tou Ng. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, pages 41–48, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[Lenat 1995] Douglas B. Lenat. CYC: A Large-Scale Investment in Knowledge Infrastructure. Communications of the ACM, vol. 38, no. 11, pages 33–38, 1995.

[Lesk 1986] Michael Lesk. Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. In Proceedings of the 1986 SIGDOC Conference, pages 24–26, New York, 1986. Association for Computing Machinery.

[Levin 1993] Beth Levin. English verb classes and alternations: A preliminary investigation. Chicago: The University of Chicago, 1993.

[Lin & Pantel 2001] Dekang Lin et Patrick Pantel. DIRT - Discovery of Inference Rules from Text. In In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 323–328, 2001.

[Lin 1998] Dekang Lin. Automatic Retrieval and Clustering of Similar Words. In Proceedings of COLING-ACL98, pages 768–774, 1998.

[Litkowski 2004a] Ken Litkowski. Automatic labeling of semantic roles. International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, 2004.

[Litkowski 2004b] Ken Litkowski. Senseval-3 task: Word Sense Disambiguation of WordNet glosses. In Rada F. Mihalcea et Phil Edmonds, editeurs, Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, pages 13–16, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[Loupy (de) 2000] Claude Loupy (de). Évaluation de l'Apport de Connaissances Linguistiques en Recherche Documentaire et Désambiguïsation Sémantique. PhD thesis, Laboratoire d'Informatique d'Avignon, 2000.

[Lund & Burgess 1996] K. Lund et C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments, and Computers, vol. 28, pages 203–208, 1996.

[Lyons 1970] John Lyons. Linguistique générale, 1968, trad. française larousse, 1970. 1970.

[M. Suchanek *et al.* 2006] Fabian M. Suchanek, Georgiana Ifrim et Gerhard Weikum. Combining linguistic and statistical analysis to extract relations from web documents. In In Proceedings of KDD 2006, page 06. KDD, 2006.

[Madsen *et al.* 2004] Rasmus Elsborg Madsen, Lars Kai Hansen et Ole Winther. Singular value decomposition and principal component analysis. Rapport technique, Informatics and Mathematical Modelling Technical University of Denmark, February 2004.

[Manandhar *et al.* 2010] Suresh Manandhar, Ioannis P. Klapaftis, Dmitriy Dligach et Sameer S. Pradhan. SemEval-2010 Task 14: Word Sense Induction & Disambiguation. In Proceedings of the 5th International Workshop on Semantic Evaluation, pages 63–68, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[Masterman 1957] Margaret Masterman. The thesaurus in syntax and semantics. Mechanical Translation, vol. 4, no. 1-2, pages 35–43, 1957.

[Mel'čuk 2010] Igor Mel'čuk. La phraséologie en langue, en dictionnaire et en TALN. In Actes de TALN 2010, Montréal, Canada, July 2010.

[Mertens 2010] Piet Mertens. Restrictions de sélection et réalisations syntagmatiques dans DICOVALENCE. Conversion vers un format utilisable en TAL. In Actes TALN 2010, Montréal, juillet 2010.

[Meurs *et al.* 2008] Marie-Jean Meurs, Frédéric Duvert, Frédéric Béchet, Fabrice Lefèvre et Renato Mori (de). Annotation en Frames Sémantiques du corpus de dialogue MEDIA. In Actes de TALN 2008 (Traitement automatique des langues naturelles), Avignon, Juin 2008. ATALA, LIA.

[Mihalcea & Faruque 2004] Rada F. Mihalcea et Ehsanul Faruque. SenseLearner: Minimally supervised Word Sense Disambiguation for all words in open text. International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, 2004.

[Mihalcea & Moldovan 2001] Rada F. Mihalcea et Dan I. Moldovan. A Highly Accurate Bootstrapping Algorithm For Word Sense Disambiguation, 2001.

[Miller *et al.* 1990] George A. Miller, R. Beckwith, Christiane. D. Fellbaum, D. Gross et K. Miller. WordNet: "An on-line lexical database". Journal of Lexicography (special issue), vol. 3, no. 4, pages 235–312, 1990.

[Miller 1995] George A. Miller. WordNet: a lexical database for English. Commun. ACM, vol. 38, no. 11, pages 39–41, 1995.

[Mitchell & Lapata 2010] Jeff Mitchell et Mirella Lapata. Composition in Distributional Models of Semantics. To appear in Cognitive Science, 2010.

[Moldovan *et al.* 2002] Dan Moldovan, A Harabagiu, Roxana Girju, Paul Morărescu, Finley Lacatusu, Adrian Novischi, Adriana Badulescu et Orest Bolohan. Lcc tools for question answering. In TREC 2002 Proceedings, 2002.

[Moldovan *et al.* 2004] Dan Moldovan, Roxana Gîrju, Marian Olteanu et Ovidiu Fortu. SVM classification of FrameNet semantic roles. In Rada F. Mihalcea et Phil Edmonds, editeurs, Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, pages 167–170, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[Moldovan *et al.* 2007] Dan I. Moldovan, Christine Clark et Moldovan Bowden. Lymba's PowerAnswer 4 in TREC 2007. In Ellen M. Voorhees et Lori P. Buckland, editeurs, TREC, volume Special Publication 500-274. National Institute of Standards and Technology (NIST), 2007.

[Moreda *et al.* 2007] Paloma Moreda, Borja Navarro et Manuel Palomar. Corpus-based semantic role approach in information retrieval. Data & Knowledge Engineering, pages 467–483, 2007.

[Moreda *et al.* 2008] Paloma Moreda, H. Llorens, E. Saquete et Manuel Palomar. The influence of Semantic Roles in QA: A comparative analysis. Procesamiento del Lenguaje Natural, pages 55–62, 2008.

[Morin & Jacquemin 2004] Emmanuel Morin et Christian Jacquemin. Automatic Acquisition and Expansion of Hypernym Links. Computers and the Humanities (CHUM), vol. 38, no. 4, pages 363–396, 2004.

[Moschitti *et al.* 2003] A. Moschitti, Paul Morărescu et Sanda M. Harabagiu. Open-domain information extraction via automatic semantic labeling. In Proceedings of FLAIRS, 2003.

[Moschitti *et al.* 2007] Alessandro Moschitti, Silvia Quarteroni, Roberto Basili et Suresh Manandhar. Exploiting Syntactic and Shallow Semantic Kernels for Question/Answer Classification. In Proceedings of the 45th Conference of the Association for Computational Linguistics (ACL), Prague, June 2007.

[Nadas *et al.* 1991] A. Nadas, D. Nahamoo, M. A. Picheny et J. Powell. An iterative 'flip-flop' approximation of the most informative split in the construction of decision trees. In ICASSP '91: Proceedings of the Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference, pages 565–568, Washington, DC, USA, 1991. IEEE Computer Society.

[Nakov & Hearst 2003] Preslav I. Nakov et Marti A. Hearst. Category-based pseudowords. In NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pages 67–69, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[Narayanan & Harabagiu 2004] Srini Narayanan et Sanda Harabagiu. Question answering based on semantic structures. In COLING '04: Proceedings of the 20th international conference on Computational Linguistics, page 693, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[Nath Jha 2004] Girish Nath Jha. The system of Panini. Language in India, vol. 4, no. 2, 2004.

[Navigli & Velardi 2003] Roberto Navigli et Paola Velardi. An Analysis of Ontology-based Query Expansion Strategies. In Proceedings of Workshop on Adaptive Text Extraction and Mining at the 14th European Conference on Machine Learning (2003), 2003.

[Ng & Lee 1996] Hwee Tou Ng et Hian Beng Lee. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In Proceedings of ACL'96, pages 40–47, 1996.

[Niles & Pease 2001] Ian Niles et Adam Pease. Towards a standard upper ontology. In FOIS '01: Proceedings of the international conference on Formal Ontology in Information Systems, pages 2–9, New York, NY, USA, 2001. ACM.

[Padó & Lapata 2005] Sebastian Padó et Mirella Lapata. Cross-lingual Bootstrapping for Semantic Lexicons: The case of FrameNet. In Proceedings of AAAI-05, pages 1087–1092, Pittsburgh, PA, 2005.

[Padó & Lapata 2007] Sebastian Padó et Mirella Lapata. Dependency-Based Construction of Semantic Space Models. Computational Linguistics, vol. 33, no. 2, pages 161–199, 2007.

[Padó & Pitel 2007] Sebastian Padó et Guillaume Pitel. Annotation précise du français en sémantique de rôles par projection cross-linguistique. In Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (communications orales), page 271, 2007.

[Palmer *et al.* 2005] Martha Palmer, Daniel Gildea et Paul Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics, 2005.

[Pantel & Lin 2002] Patrick Pantel et Dekang Lin. Discovering word senses from text. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 613–619, New York, NY, USA, 2002. ACM Special Interest Group on Knowledge Discovery in Data, ACM Press, ISBN:1-58113-567-X.

[Pantel & Pennacchiotti 2006] Patrick Pantel et Marco Pennacchiotti. Espresso: leveraging generic patterns for automatically harvesting semantic relations. In ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pages 113–120, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[Partee 1984] B. Partee. Compositionality, pages 281–311. Dordrecht: Foris, 1984.

[Patterson & Berthold 2001] David Patterson et Michael R. Berthold. Finding clusters in parallel universes. In Systems, Man, and Cybernetics, 2001 IEEE International Conference on, pages 123–128, 2001.

[Patwardhan *et al.* 2003] Siddharth Patwardhan, Satanjeev Banerjee et Ted Pedersen. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, pages 241–257, Mexico City, Mexico, February 2003.

[Peat & Willett 1991] Helen J. Peat et Peter Willett. The limitations of term co-occurrence data for query expansion in document retrieval systems. Journal of the American Society for Information Science, vol. 42, pages 378–383, 1991.

[Pedersen *et al.* 2004] Ted Pedersen, Siddharth Patwardhan et Jason Michelizzi. WordNet::Similarity - Measuring the Relatedness of Concepts. In Proceedings of NAACL-04, pages 1024–1025, 2004.

[Peirsman & Padó 2010] Yves Peirsman et Sebastian Padó. Cross-lingual Induction of Selectional Preferences with Bilingual Vector Spaces. In Proceedings of NAACL/HLT, Los Angeles, CA, 2010.

[Pennacchiotti *et al.* 2008] Marco Pennacchiotti, Diego Cao (de), Roberto Basili, Danilo Croce et Michael Roth. Automatic induction of FrameNet lexical units. In Proceedings of Empirical Methods in Natural Language Processing (EMNLP), Honolulu, Hawaii, USA, 2008.

[Pizzato & Mollá-Aliod 2005] Luiz Augusto Sangoi Pizzato et Diego Mollá-Aliod. Extracting Exact Answers using a Meta Question answering System. In Proceedings of the Australasian Language Technology Workshop 2005 (ALTW05), Sidney, Australia, December 2005.

[Pizzato & Mollá 2008] Luiz Augusto Pizzato et Diego Mollá. Indexing on semantic roles for question answering. In IRQA '08: Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering, pages 74–81, Morristown, NJ, USA, 2008. Association for Computational Linguistics.

[Ploux & Victorri 1998] Sabine Ploux et Bernard Victorri. Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. Traitement automatique des langues, vol. 39, no. 1, 1998.

[Pradhan *et al.* 2007] Sameer S. Pradhan, Edward Loper, Dmitriy Dligach et Martha Palmer. SemEval-2007 task 17: English lexical sample, SRL and all words. In SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations, pages 87–92, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

[Pradhan *et al.* 2008] Sameer S. Pradhan, Wayne Ward et James H. Martin. Towards robust semantic role labeling. Computational Linguistics, vol. 34, no. 2, pages 289–310, 2008.

[Preiss & Stevenson 2004] Judita Preiss et Mark Stevenson. Editorial - Introduction to the special issue on word sense disambiguation. Computer Speech and Language, vol. 18, pages 201–207, 2004.

[Purandare 2004] Amruta Purandare. Word sense discrimination by clustering similarity contexts. Master Thesis, 2004.

[Qiu & Frei 1993] Yonggang Qiu et Hans-Peter Frei. Concept based query expansion. In SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pages 160–169, New York, NY, USA, 1993. ACM.

[Rapp 2004] Reinhard Rapp. A practical solution to the problem of automatic word sense induction. In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, page 26, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[Rasheed & Shah 2002] Z. Rasheed et M. Shah. Movie genre classification by exploiting audio-visual features of previews. In Pattern Recognition, 2002. Proceedings. 16th International Conference on, volume 2, pages 1086–1089 vol.2, 2002.

[Ravichandran *et al.* 2005] Deepak Ravichandran, Patrick Pantel et Eduard Hovy. Randomized algorithms and NLP: using locality sensitive hash function for high speed noun clustering. In ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 622–629, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[Reed & Lenat 2002] Stephen L. Reed et Douglas B. Lenat. Mapping Ontologies into Cyc, 2002.

[Rieger & Small 1979] Chuck Rieger et Steven Small. Word expert parsing. In Proceedings of the 6th International Joint Conference on Artificial Intelligence (IJCAI), pages 723–728, 1979.

[Rocchio 1971] R. Rocchio, editeur. Relevance feedback in information retrieval. Prentice-Hall (Englewood Cliffs, N.J), 1971.

[Rosen 1984] C. Rosen. Studies in relationnal grammar 2, chapitre The interface between semantic roles and initial grammatical relations, pages 38–77. IL : University of Chicago Press, Chicago, 1984.

[Rosenberg & Hirschberg 2007] Andrew Rosenberg et Julia Hirschberg. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 410–420, 2007.

[Ruiz-Casado *et al.* 2006] Maria Ruiz-Casado, Enrique Alfonseca et Pablo Castells. Automatising the learning of lexical patterns: An application to the enrichment of WordNet by extracting semantic relationships from Wikipedia. Data & Knowledge Engineering, 2006.

[Russell & Norvig 2003] Stuart Russell et Peter Norvig. Artificial intelligence: A modern approach. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition édition, 2003.

[Sagot & Fišer 2008] Benoît Sagot et Darja Fišer. Combining Multiple Resources to Build Reliable Wordnets. In Proceedings of the 11th international conference on Text, Speech and Dialogue, pages 61 – 68, Brno, Czech Republic, 2008.

[Sahlgren 2006] Magnus Sahlgren. The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. PhD thesis, Department of Linguistics, Stockholm University, 2006.

[Saint-Dizier 2006] Patrick Saint-Dizier. Rôles thématiques. In Danièle Godard, Laurent Roussarie et Francis Corblin, editeurs, Sémanticlopédie: dictionnaire de sémantique. GDR Sémantique & Modélisation, CNRS, http://www.semantique-gdr.net/dico/, 2006.

[Salton & Buckley 1990] Gerard Salton et Chris Buckley. Improving Retrieval Performance by Relevance Feedback. Journal of the ASIS, vol. 41, no. 4, pages 288–297, 1990.

[Salton *et al.* 1975] Gerard Salton, A. Wong et C. S. Yang. A Vector Space Model for Automatic Indexing. Communications of the ACM, vol. 18, no. 11, pages 613–620, 1975.

[Salvo Braz (de) *et al.* 2005] Rodrigo Salvo Braz (de), Roxana Girju, Vasin Punyakanok, Dan Roth et Mark Sammons. Knowledge Representation for Semantic Entailment and Question-Answering. In IJCAI'05: Workshop on Knowledge and Reasoning for Question Answering, 2005.

[Sanderson 2000] Mark Sanderson. Retrieving with good sense. Information Retrieval, vol. 2, no. 1, pages 49–69, 2000.

[Sarjant *et al.* 2009] Samuel Sarjant, Catherine Legg, Michael Robinson et Olena Medelyan. "All You Can Eat" Ontology-Building: Feeding Wikipedia to Cyc. Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on, vol. 1, pages 341–348, 2009.

[Saussure (de) 1916] Ferdinand Saussure (de). Cours de linguistique générale, 1916.

[Schuler 2005] Karin Kipper Schuler. VerbNet: A broad-coverage, comprehensive verb lexicon. Univ. of Pennsylvania-Electronic Dissertations, 2005.

[Schütze 1998] Hinrich Schütze. Automatic word sense discrimination. Computational Linguistics, vol. 24, no. 1, pages 97–123, 1998.

[Schwab *et al.* 2007] Didier Schwab, Lim Lian Tze et Mathieu Lafourcade. Les vecteurs conceptuels, un outil complémentaire aux réseaux lexicaux. In Actes de TALN 07, 2007.

[Schütze 1992] Hinrich Schütze. Dimensions of Meaning. In Proceedings of Supercomputing '92, 1992.

[Segond *et al.* 1997] Frédérique Segond, Anne Schiller, Gregory Grefenstette et Jean-Pierre Chanod. An Experiment in Semantic Tagging using Hidden Markov Model Tagging. In Piek Vossen, Geert Adriaens, Nicoletta Calzolari et Antonio Sanfilippo amd Yorick Wilks, editeurs, Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, pages 78–81. Association for Computational Linguistics, New Brunswick, New Jersey, 1997.

[Segond 2000] Frédérique Segond. Framework and results for French. Computers and the Humanities, Special Issue on SENSEVAL, vol. 34, no. 1-2, 2000.

[Shafer 1976] Glenn Shafer. A mathematical theory of evidence. Princeton University Press, 1976.

[Shen & Lapata 2007] Dan Shen et Mirella Lapata. Using semantic roles to improve question answering. In Proceedings of EMNLP-CoNLL, 2007.

[Shi & Mihalcea 2004] Lei Shi et Rada F. Mihalcea. An Algorithm for Open Text Semantic Parsing. In In Proceedings of the ROMAND 2004 workshop on "Robust Methods in Analysis of Natural language Data, 2004.

[Snow 2006] Rion Snow. Semantic taxonomy induction from heterogenous evidence. In In Proceedings of COLING/ACL 2006, pages 801–808, 2006.

[Snyder & Palmer 2004] Benjamin Snyder et Martha Palmer. The English all-words task. International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, 2004.

[Sowa 1984] John F. Sowa. Conceptual structures : processing in mind and machine. Addison-Wesley, Reading, Massachusetts, 1984.

[Sowa 2000] John F. Sowa. Knowledge representation: logical, philosophical and computational foundations. Brooks/Cole Publishing Co., Pacific Grove, CA, USA, 2000.

[Sparck Jones *et al.* 2000] K. Sparck Jones, S. Walker et S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. Inf. Process. Manage., vol. 36, no. 6, pages 779–808, 2000.

[Stenchikova *et al.* 2006] Svetlana Stenchikova, Dilek Hakkani-Tur et Gokan Tur. QASR: Question Answering Using Semantic Roles for Speech Interface. In Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP-Interspeech 2006), pages 1185–1188, Pittsburgh, Pennsylvania, 2006.

[Stevenson & Wilks 2001] Mark Stevenson et Yorick Wilks. The interaction of knowledge sources in word sense disambiguation. Computational Linguistics, vol. 27, no. 3, pages 321 – 349, September 2001.

[Strehl & Ghosh 2002] Alexander Strehl et Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res., vol. 3, pages 583–617, 2002.

[Strzalkowski 1994] T. Strzalkowski. Building a lexical domain map from text corpora. In Proceedings of the 15th International Conference on Computational Linguistics, COLING'94, pages 604–610, 1994.

[Suchanek *et al.* 2007] Fabian M. Suchanek, Gjergji Kasneci et Gerhard Weikum. Yago: A Core of Semantic Knowledge. In 16th international World Wide Web conference (WWW 2007), New York, NY, USA, 2007. ACM Press.

[Sun *et al.* 2005] R. X. Sun, J. J. Jiang, Y. F. Tan, H. Cui, T. S. Chua et M. Y. Kan. Using syntactic and semantic relation analysis in question answering. In Proceedings of the TREC, 2005.

[Surdeanu *et al.* 2003] Mihai Surdeanu, Sanda Harabagiu, John Williams et Paul Aarseth. Using predicate-argument structures for information extraction. In Proceedings of ACL, 2003.

[Surdeanu *et al.* 2008] Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez et Joakim Nivre. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In CoNLL '08: Proceedings of the Twelfth Conference on Computational Natural Language Learning, pages 159–177, Morristown, NJ, USA, 2008. Association for Computational Linguistics.

[Swier & Stevenson 2004] Robert S. Swier et Suzanne Stevenson. Unsupervised Semantic Role Labeling. In Proceedings of EMNLP 2004, pages 95–102, Barcelona, Spain, 2004.

[Tenny 1994] C. L. Tenny. Aspectual roles and the syntax-semantics interface. Kluwer Academic Publishers, 1994.

[Thompson *et al.* 2004] Cynthia A. Thompson, Siddharth Patwardhan et Carolin Arnold. Generative models for semantic role labeling. In Rada F. Mihalcea et Phil Edmonds, editeurs, Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, pages 235–238, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[Thompson 2004] Cynthia A. Thompson. Semi-supervised Semantic Role Labeling, 2004.

[Thorndike 1941] Edward Lee Thorndike. Thorndike century beginning dictionary. Scott, Foresman and Co, 1941.

[Tomuro *et al.* 2007] Noriko Tomuro, Steven L. Lytinen, Kyoko Kanzaki et Hitoshi Isahara. Clustering Using Feature Domain Similarity to Discover Word Senses for Adjectives. In Proceedings of ICSC, 2007.

[Tonelli & Pianta 2008] Sara Tonelli et Emanuele Pianta. Frame information transfer from English to Italian. Proceedings of LREC-2008, 2008.

[Topchy *et al.* 2004] Alexander P. Topchy, Martin H. C. Law, Anil K. Jain et Ana L. Fred. Analysis of Consensus Partition in Cluster Ensemble. In ICDM '04: Proceedings of the Fourth IEEE International Conference on Data Mining, pages 225–232, Washington, DC, USA, 2004. IEEE Computer Society.

[Topchy *et al.* 2005] Alexander P. Topchy, Anil K. Jain et William F. Punch. Clustering ensembles: models of consensus and weak partitions. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 27, no. 12, pages 1866–1881, October 2005.

[Toutanova *et al.* 2005] Kristina Toutanova, Aria Haghighi et Christopher D. Manning. Joint learning improves semantic role labeling. In Proceedings of ACL 2005, 2005.

[Tratz *et al.* 2007] Stephen Tratz, Antonio Sanfilippo, Michelle Gregory, Alan Chappell, Christian Posse et Paul Whitney. PNNL: a supervised maximum entropy approach to word sense disambiguation. In SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations, pages 264–267, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

[Turney 2001] Peter D. Turney. Mining the Web for Synonyms: PMI–IR versus LSA on TOEFL . Lecture Notes in Computer Science, vol. 2167, pages 491–502, 2001.

[Van Der Plas & Tiedemann 2006] Lonneke Van Der Plas et Jörg Tiedemann. Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity. In Proceedings of the COLING/ACL 2006, 2006.

[Vasilescu & Langlais 2004] Florentina Vasilescu et Philippe Langlais. Désambiguïsation de corpus monolingues par des approches de type Lesk. In Actes de TALN 04, 2004.

[Vechtomova & Karamuftuoglu 2007] Olga Vechtomova et Murat Karamuftuoglu. Query expansion with terms selected using lexical cohesion analysis of documents. Information Processing and Management: an International Journal, vol. 43, no. 4, pages 849–865, July 2007.

[Vechtomova *et al.* 2003] Olga Vechtomova, Stephen Robertson et Susan Jones. Query expansion with long-span collocates. Information Retrieval, vol. 6, no. 2, pages 251–273, 2003.

[Vechtomova *et al.* 2006] Olga Vechtomova, Murat Karamuftuoglu et S.E. Robertson. On document relevance and lexical cohesion between query terms. Information Processing and Management, vol. 42, no. 5, pages 1230–1247, 2006.

[Velldal 2005] Erik Velldal. A Fuzzy clustering approach to word sense discrimination. In Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, Copenhagen, Denmark, 2005.

[Venant 2007] Fabienne Venant. Utiliser des classes de sélection distributionnelle pour désambiguïser les adjectifs. In Actes de TALN 07, 2007.

[Véronis & Ide 1990] Jean Véronis et Nancy M. Ide. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In Proceedings of the 13th conference on Computational linguistics, pages 389–394, Morristown, NJ, USA, 1990. Association for Computational Linguistics.

[Véronis 2001] Jean Véronis. Sense tagging: does it make sense ? In In The Corpus Linguistics Conference, Lancaster, UK, 2001.

[Véronis 2003] Jean Véronis. Cartographie lexicale pour la recherche d'information. In Béatrice Daille, editeur, Actes de TALN 2003 (Traitement automatique des langues naturelles), pages 265–274, Batz-sur-mer, Juin 2003. ATALA, IRIN.

[Véronis 2004] Jean Véronis. HyperLex: lexical cartography for information retrieval. Computer Speech and Language, vol. 18, pages 223–252, 2004.

[Victorri & Fuchs 1996] Bernard Victorri et Christian Fuchs. La polysémie, construction dynamique du sens. Hermès, Paris, 1996.

[Vilnat 2005] Anne Vilnat. Dialogue et analyse de phrases. Mémoire d'Habilitation à diriger des recherches, Décembre 2005.

[Voorhees 1993] Ellen M. Voorhees. Using WordNet to disambiguate word senses for text retrieval. In SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pages 171–180, New York, NY, USA, 1993. ACM.

[Voorhees 1994] Ellen M. Voorhees. Query expansion using lexical-semantic relations. In SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pages 61–69, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[Vossen 1998] Piek Vossen, editeur. Eurowordnet: A multilingual database with lexical semantic networks. Kluwer Academic Publishers, 1998.

[Walkowiak 2010] Adrien Walkowiak. Caractérisation d'un espace sémantique bimodal texte-image. Rapport de stage, 2010.

[Weaver 1949] Warren Weaver. Machine translation of languages: Fourteen essays, chapitre Translation, pages 15–23. The Technology Press of the Massachussetts Institute of TechnologyJohn Wiley (New York)Clapham & Hall, London, 1949.

[Widdows 2008] Dominic Widdows. Semantic Vector Products: Some Initial Investigations. In Second AAAI Symposium on Quantum Interaction, Oxford, March 2008.

[Wilks *et al.* 1990] Yorick Wilks, Dan Fass, Cheng-Ming Guo, James E. MacDonald, Tony Plate et Brian A. Slator Slator. Providing machine tractable dictionary tools. Machine Translation, vol. 5, no. 2, pages 99–154, June 1990.

[Wiswedel & Berthold 2007] Bernd Wiswedel et Michael R. Berthold. Fuzzy clustering in parallel universes. International Journal of Approximate Reasoning, vol. 45, no. 3, pages 439–454, August 2007.

[Wong 1982] Douglas Wong. On the unification of language comprehension with problem solving. PhD thesis, Brown University, Providence, RI, USA, 1982.

[Xu & Croft 2000] Jinxi Xu et W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. ACM Trans. Inf. Syst., vol. 18, no. 1, pages 79–112, 2000.

[Xue & Palmer 2004] Nianwen Xue et Martha Palmer. Calibrating features for semantic role labeling. In In Proceedings of EMNLP 2004, pages 88–94, 2004.

[Yager 1988] R. R. Yager. On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Transactions on Systems, Man, and Cybernetics, vol. 18, page 183–190, 1988.

[Yang & Powers 2006] Dongqiang Yang et David M. W. Powers. Word sense disambiguation using lexical cohesion in the context. In Proceedings of the COLING/ACL on Main conference poster sessions, pages 929–936, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[Yarowsky 1992] David Yarowsky. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In Proceedings of COLING 14, pages 454–460, 1992.

[Yarowsky 1993] David Yarowsky. One sense per collocation. In Proceedings of ARPA Human Language Technology Workshop, pages 266–271, Princeton, NJ, USA, 1993.

[Yarowsky 1995] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of ACL 33, pages 189–196, 1995.

[Zhao *et al.* 2006] Feng Charlie Zhao, Yugyung Lee et Deep Mehdi. Experiments with Query Expansion at TREC 2006 Legal Track. In The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings, 2006.

[Zipf 1949] George Kingsley Zipf. Human behavior and the principle of least effort: An introduction to human ecology. Hafner Pub. Co; Facsim. of 1949 ed edition (1972), 1949.

# Annexes

# Annexe A

# Glossaire

**Classification automatique**

On appelle classification automatique la catégorisation algorithmique d'objets. Celle-ci consiste à attribuer une classe ou catégorie à chaque objet (ou individu) à classer, en se basant sur des données statistiques. Cela fait couramment appel à l'apprentissage automatique et est largement utilisé en reconnaissance de formes.
Source : *Wikipedia* .

**Classification k-nearest neighbours**

En intelligence artificielle, la méthode des $k$ plus proches voisins est une méthode d'apprentissage supervisé.

Dans ce cadre, on dispose d'une base de données d'apprentissage constituée de $N$ couples « entrée-sortie ». Pour estimer la sortie associée à une nouvelle entrée $x$, la méthode des $k$ plus proches voisins consiste à prendre en compte (de façon identique) les $k$ échantillons d'apprentissage dont l'entrée est la plus proche de la nouvelle entrée $x$, selon une distance à définir.

Par exemple, dans un problème de classification, on retiendra la classe la plus représentée parmi les $k$ sorties associées aux $k$ entrées les plus proches de la nouvelle entrée $x$.

Source : *Wikipedia* .

**Clustering**

Le partitionnement de données (*data clustering* en anglais) est une des méthodes Statistiques d'analyse des données. Elle vise à diviser un ensemble de données en différents « paquets » homogènes, en ce sens que les données de chaque sous-ensemble partagent des caractéristiques communes, qui correspondent le plus souvent à des critères de proximité (similarité informatique) que l'on définit en introduisant des mesures et classes de distance entre objets [1].

Pour obtenir un bon partitionnement, il convient d'à la fois :
* minimiser l'inertie intra-classe pour obtenir des grappes (*clusters* en anglais) les plus homogènes possibles.
* maximiser l'inertie inter-classe afin d'obtenir des sous-ensembles bien différenciés.

Source : *Wikipedia* .

**Expressions multimots**

Une expression multimots est un agencement de mots particulier employé à une fréquence significative et dont le sens est fixé.

Elle peut être compositionnelle, lorsque le sens du tout est constitué par le sens des parties, comme c'est le cas pour l'exemple *ticket de métro*, ou bien non compositionnelle lorsqu'il s'agit d'une expression idiomatique (dite aussi sémantiquement opaque), comme c'est le cas pour l'exemple *casser les pieds*. .

**Indice de Jaccard**

L'indice et la distance de Jaccard sont deux métriques utilisées en statistiques pour comparer la similarité et la diversité entre des échantillons. Elles sont nommées d'après le botaniste suisse Paul Jaccard.

L'indice de Jaccard (ou coefficient de Jaccard) est le rapport entre la cardinalité (la taille) de l'intersection des ensembles considérés et la cardinalité de l'union des ensembles. Il permet d'évaluer la similarité entre les ensembles. Soit deux ensembles A et B, l'indice est :

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

L'extension à n ensembles est triviale :

$$J(S_1, S_2, ..., S_n) = \frac{|S_1 \cap S_2 \cap ... \cap S_n|}{|S_1 \cup S_2 \cup ... \cup S_n|}$$

La distance de Jaccard mesure la dissimilarité entre les ensembles. Elle consiste simplement à soustraire l'indice de Jaccard à 1.

$$J_\delta(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

De la même manière que pour l'indice, la généralisation devient :

$$J_\delta(S_1, S_2, ..., S_n) = 1 - J(S_1, S_2, ..., S_n) = \frac{|S_1 \cup S_2 \cup ... \cup S_n| - |S_1 \cap S_2 \cap ... \cap S_n|}{|S_1 \cup S_2 \cup ... \cup S_n|}$$

Source : *Wikipedia* .

**Ordered Weighted Averaging (OWA)**

Un mapping

$$F : [0,1]^n \to [0,1]$$

est appelé opérateur OWA de dimension $n$ s'il est associé avec un vecteur de poids $w = [w_1, ..., w_n]$ tel que :

$$1)\ w_i \in [0,1]$$

$$2)\ \sum_i w_i = 1$$

$$3)\ F(a_1, ..., a_n) = \sum_{i=1}^{n} w_i b_i$$

avec $b_i$ le i-ième plus grand élément de la collection $a_1, ..., a_n$.
Source : [Le *et al.* 2007] .

**Théorie de l'évidence de Dempster-Shafer**

Soit $\Theta$ l'univers et $P(\Theta)$ l'ensemble des parties de l'ensemble $\Theta$ (*(*Power set)).
L'assignation basique de probabilités (BPA) (autrement appelée masse) est une fonction $m$ : $P(\Theta) \to [O, 1]$ telle que :

$$m(\varnothing) = 0 \text{ et } \sum_{A \in P(\Theta)} m(A) = 1$$

La quantité $m(A)$ peut être interprétée comme la mesure de la croyance qui est accordée à A en fonction des connaissances disponibles. Un sous-ensemble $A \in 2^\Theta$ avec $m(A) > 0$ est appelé *élément focal* de $m$.
Deux fonctions dérivées de la BPA $m$ sont les fonctions de *croyance (Bel)* et de *plausibilité (Pl)*, bornes respectivement inférieure et supérieure de la probabilité exacte de l'événement A.

$$Bel_m(A) = \sum_{\varnothing \neq B \subseteq A} m(B)$$

$$Pl_m(A) = \sum B \cap A = \varnothing m(B)$$

*Fonction de discounting.* Si une source d'information a un degré de confiance $\alpha$, on peut employer la règle de discount suivante :

$$m\alpha(A) = \alpha m(A), \forall A \subset \Theta$$

$$m\alpha(\Theta) = (1 - \alpha) + \alpha m(\Theta)$$

*Règle de combinaison de Dempster.* Afin de combiner l'information provenant de différentes sources, Dempster définit ainsi la combinaison de ces information dans la BPA résultante :

$$(m1 \bigoplus m2)(\varnothing) = 0,$$

$$(m1 \bigoplus m2)(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C)$$

$$\text{avec } \kappa = \sum_{B \cap C = \varnothing} m_1(B) m_2(C)$$

Cette combinaison orthogonale ne peut être appliquée que si $\kappa > 1$.

Source : [Le *et al.* 2007] .

**W3C**

Le World Wide Web Consortium, abrégé par le sigle W3C, est un organisme de standardisation à but non-lucratif, fondé en octobre 1994 comme un consortium chargé de promouvoir la compatibilité des technologies du World Wide Web telles que HTML, XHTML, XML, RDF, CSS, PNG, SVG et SOAP. Le W3C n'émet pas des normes au sens européen, mais des recommandations à valeur de standards industriels.

Source : *Wikipedia* .

**Web Sémantique**

Le Web sémantique désigne un ensemble de technologies visant à rendre le contenu des ressources du World Wide Web accessible et utilisable par les programmes et agents logiciels, grâce à un système de métadonnées formelles, utilisant notamment la famille de langages développés par le W3C.

Source : *Wikipedia* .

# Annexe B

# Liste des publications

Mouton C. Induction de sens de mots à partir de multiples espaces sémantiques. Actes de RECITAL 2009.

Mouton C., Pitel G., de Chalendar G., Vilnat, A. Unsupervised Word Induction from Multiple Semantic Spaces. Proceedings of RANLP 2009.

Mouton C., Richert B. de Chalendar G.Traduction de FrameNet par dictionnaires bilingues avec évaluation sur la paire anglais-français. Actes de MajecSTIC 2009.

Mouton C., de Chalendar G., Richert B. FrameNet Translation Using Bilingual Dictionaries with Evaluation on the French-English Pair. Proceedings of LREC 2010.

Mouton C., de Chalendar G. JAWS : Just Another WordNet Subset. Actes de TALN 2010.