Construction automatique de classes de sélection distributionnelle

Guillaume Jacquet, Fabienne Venant

Construction automatique de classes de sélection distributionnelle

Guillaume Jacquet, Fabienne Venant

LaTTICe – CNRS UMR 8094
Langues, Textes, Traitements Informatiques et Cognition
Ecole Normale Supérieure
1 rue Maurice Arnoux
F-92120 Montrouge
guillaume.jacquet@ens.fr
fabienne.venant@ens.fr

Mots-clés : classes de sélection distributionnelle, espace distributionnel, désambiguïsation, corpus, contexte
Keywords : semantic classes, distributional space, disambiguation, corpus, context

Résumé

Cette étude se place dans le cadre général de la désambiguïsation automatique du sens d'un verbe dans un énoncé donné. Notre méthode de désambiguïsation prend en compte la construction du verbe, c'est-à-dire l'influence des éléments lexicaux et syntaxiques présents dans l'énoncé (cotexte). Nous cherchons maintenant à finaliser cette méthode en tenant compte des caractéristiques sémantiques du cotexte. Pour ce faire nous associons au corpus un espace distributionnel continu dans lequel nous construisons et visualisons des classes distributionnelles. La singularité de ces classes est qu'elles sont calculées à la volée. Elles dépendent donc non seulement du corpus mais aussi du contexte étudié. Nous présentons ici notre méthode de calcul de classes ainsi que les premiers résultats obtenus.

Abstract

This study is placed within the general framework of the automatic verb sense disambiguation. To assign a meaning to a verb, we take into account the construction of the verb, i.e. the other lexical and syntactic units within the utterance (co-text). We now seek to finalize our method by taking into account the semantic features of this co-text. We associate with the corpus a continuous distributional space in which we build and visualize distributional classes. The singularity of these classes is that they are computed "on line" for disambiguate a given context in the given corpus. They thus depend not only on the corpus but also on the studied context. We present here our method of computation of classes and first results obtained.

1 Introduction

L'objet de cette étude est la désambiguïsation automatique du sens d'un mot en contexte. Elle s'inscrit dans le cadre théorique de la construction dynamique du sens, proposé par Victorri et Fuchs (1996) : on associe à chaque unité polysémique un espace sémantique. Le sens de l'unité dans un énoncé donné correspond à une région plus ou moins étendue de cet espace, déterminée par l'interaction dynamique de toutes les unités présentes dans l'énoncé. Ploux et Victorri ont développé un logiciel, Visusyn, permettant de construire automatiquement l'espace sémantique associé à une unité lexicale à partir d'un dictionnaire de synonymes (Ploux, Victorri, 1998). Ce logiciel a ensuite été étendu de façon à pouvoir prendre en compte des données distributionnelles issues d'un corpus dans une tâche de désambiguïsation automatique (François et al, 2003 ; Venant, 2004). Une réflexion sur le rôle de la syntaxe dans la construction du sens, s'appuyant notamment sur les travaux de Goldberg (1995) et Kay (2000), a permis d'enrichir le logiciel d'un nouveau module. Considérant que les constructions syntaxiques sont porteuses d'un sens intrinsèque et qu'à ce titre elles font partie intégrante du cotexte, ce module permet d'associer à chaque construction syntaxique une zone dans l'espace sémantique de l'unité étudiée (Jacquet, 2004). Nous cherchons maintenant à finaliser notre méthode de désambiguïsation en tenant compte des caractéristiques sémantiques du cotexte lexical. Il s'agit d'associer une zone de l'espace sémantique non plus à chaque unité rencontrée en contexte mais à des classes d'unités. Ceci permettra de traiter des noms propres ou des cooccurrences rares dans le corpus. Par exemple, le sens de *jouer du luth* sera associé à *jouer de* (*le luth, la guitare, le piano,* ….). On pourra alors, dans l'espace sémantique du verbe *jouer*, définir une zone correspondant à la classe (*luth, guitare, piano,*…), distincte de celle associée à une autre classe comme (*charme, prestance, influence*, …) pour *jouer de son charme*. Ces classes sont déterminées automatiquement à partir d'un corpus. Leur singularité est qu'elles sont dépendantes de l'énoncé étudié. Nous voulons construire non pas une ontologie de la langue française mais des classes distributionnelles avec pertinence d'emploi. Pour ce faire nous associons au corpus un espace distributionnel continu dans lequel nous construisons et visualisons les classes de sélection distributionnelle associées au contexte étudié.

2 Des classes de sélection distributionnelle

La technique que nous utilisons s'inscrit dans le cadre bien connu de l'analyse distributionnelle « à la Harris ». Elle est exploitée depuis longtemps dans la communauté du TAL pour la construction de bases de connaissances ou de ressources terminologiques à partir de textes (Frérot, 2003 ; Habert et Nazarenko, 1996 ; Fleury, 1998 ; Aussenac-Gilles et al, 2000, Pantel et Lin, 2001 ; …). Notre méthode est entièrement automatique. Elle ne fait appel à aucune modélisation préalable de connaissances sémantiques sur le corpus et utilise des rapprochements de mots sur la base de contextes syntaxiques partagés. En tout cela elle se rapproche des travaux de Greffenstette (1994). Les contextes nous sont fournis par l'analyseur Syntex (Bourigault et Fabre, 2000). Comme le précise D. Bourigault : « Là où G. Greffenstette se contente volontairement d'une analyse syntaxique relativement rudimentaire, réalisée par l'analyseur Sextan, nous avons fait le choix d'une analyse, certes encore partielle, mais plus large et plus précise, réalisée par Syntex. De ce fait, les procédures statistiques d'analyse distributionnelle de Greffenstette ne concernent que des mots simples, alors que nous pouvons prendre en compte des entités complexes (contextes ou termes) », cela nous permet de prendre en compte des distinctions plus fines, de créer des classes plus riches en information sémantique et donc plus efficaces dans leur apport à la désambiguïsation automatique. Notre travail est à rapprocher de celui de Habert et al (2004). Nous travaillons nous aussi à partir des rapports de dépendance syntaxique élémentaire entre un contexte et les

mots pleins qu'il régit ou qui le régissent et nous considérons les mots comme des points dans l'espace à n dimensions des contextes (que nous appelons l'espace distributionnel). Nous poursuivons cependant des objectifs différents. Nous ne cherchons pas à créer des classes de mots ayant le même sens mais des classes de mots dont le comportement sémantique influence de la même façon un contexte donné. Autrement dit si nous voulons trouver la classe de *luth (guitare, piano,…)* ce n'est pas pour caractériser le sens de *luth* mais pour désambiguïser *jouer* dans *jouer du luth.* Nous ne cherchons pas non plus à « faire parler le corpus dans sa globalité » comme le font Aussenac-Gilles et al (2000). Les classes qu'ils construisent se constituent en navigant autour d'éléments saillants ou prototypiques et leur permettent d'obtenir une image sémantique du corpus. Nous nous intéressons au contraire à des mots relativement peu fréquents, et qui ne représentent donc pas une ligne de force du corpus, pour rechercher dans leur classe sémantique des mots plus fréquents et dont l'apport à la désambiguïsation automatique sera plus pertinent. Certes les classes obtenues rendent compte de l'information sémantique présente dans le corpus mais de façon mouvante (Habert et al, 1999). Chaque interrogation concerne un contexte et un mot différents et donne lieu à des regroupements différents au sein de l'espace distributionnel. Nos classes s'apparentent plutôt aux classes d'objet décrites par Gaston Gross (2004) : « tout changement de sens d'un prédicat est corrélé à un changement de son schéma d'arguments. Soit la phrase *Vous suivrez ce chemin*. Si on remplace l'objet *chemin* par des substantifs comme *route, rue, voie, sentier* le verbe *suivre* garde le même sens. On regroupera ces mots sous le terme générique de <voies>. Si en revanche, on remplace le mot *chemin* par *cours*, alors on a affaire à un autre emploi et le substantif *cours* peut être remplacé par *séminaire, stage, formation, cycle d'étude*, etc., qu'on rangera sous le classifieur <enseignement> ». Nous partageons avec Gross l'idée que « la mise au point du sens exige que l'on soit à même de préciser la nature sémantique des arguments que prend un emploi donné de prédicat » mais la différence entre les classes que nous cherchons et les classes d'objets de Gross c'est que nous ne cherchons pas à établir des classes en langue. Nos classes dépendent du contexte et surtout du corpus étudié. Gross cherche à créer des classes pouvant figurer dans un dictionnaire, c'est à dire calculées une fois pour toutes sur le lexique et indépendantes du corpus étudié. Nous proposons quelque chose de plus souple. Nos classes sont calculées en ligne pour désambiguïser un contexte dans un corpus donné. Elles ne sont valables que pour ce contexte même s'il peut y avoir des recouvrements. Elles ne sont pas nécessairement générales ni référentiables par un classifieur conceptuel comme <enseignement>. Elles caractérisent un comportement sémantique au sein d'un corpus donné plutôt qu'une notion et ne sont absolument pas hiérarchisables. L'intérêt de travailler à partir d'un contexte particulier est de limiter le nombre d'éléments à classer. Lorsqu'on étudie par exemple les compléments d'un verbe donné, on ne cherche pas à classer tous les noms de la langue française mais seulement les noms pertinents dans le contexte de ce verbe. Les classes sont obtenues plus facilement et sont plus significatives que des classes construites sur la globalité du lexique.

3 Des classes de sélection distributionnelle : pourquoi ?

L'algorithme utilisé dans Visusyn repose sur l'analyse d'un graphe de synonymie. Dans ce cadre on considère qu'un sens possible pour un mot est donné par une clique de synonymes de ce mot, c'est à dire un ensemble de synonymes du mot, tous synonymes entre eux, le plus grand possible. Au stade actuel de son développement, Visusyn est capable, étant donné une construction verbale, de déterminer le sens le plus probablement pris par le verbe dans cette construction. Considérons par exemple les énoncés suivants :

1) *jouer la fille sérieuse*
2) *jouer avec sa fille*

En considérant les têtes nominales de compléments (*fille*) d'une part, et la construction syntaxique (V+SN / V+SP (avec +SN)) d'autre part, Visusyn calcule que le sens le plus probable de *jouer* est *interpréter, incarner* dans l'énoncé (1) et *s'amuser, plaisanter* dans (2). Ce modèle est opérationnel et en cours d'évaluation (Jacquet, à paraître) mais on sait d'ores et déjà qu'il échoue sur des énoncés du type :

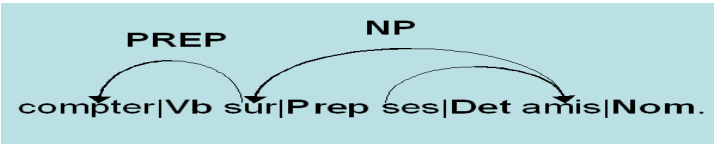
3) Jouer du luth
4) Jouer à Wimbledon

On se heurte ici à un double problème. Nous avons à faire à des compléments très peu représentés dans le corpus de référence. Or notre calcul repose sur l'utilisation des fréquences de cooccurrence de *luth* ou *Wimbledon* avec chacun des synonymes de *jouer*. Si ces fréquences sont trop faibles, le résultat du calcul est peu fiable. La première idée a été de remplacer *luth* par l'ensemble de ses synonymes, et de calculer leurs fréquences de cooccurrence avec chaque synonyme de *jouer*. Le problème est que les synonymes de *luth* sont trop peu nombreux et trop peu fréquents dans le corpus pour que le calcul soit efficace. Quand à *Wimbledon*, comme la plupart des noms propres, il ne possède aucun synonyme. Nous avons donc cherché à pallier ce manque d'informations quantitatives en fournissant à notre système des informations sémantiques sur les mots en question. Si nous pouvons associer à *luth* un ensemble de mots représentatifs des instruments de musique (*luth, guitare, piano, violon, etc*), nous retombons alors sur des énoncés interprétables par Visusyn et nous sortons de l'impasse. L'idée n'est certes pas nouvelle mais l'originalité de notre travail réside dans le fait que les classes que nous voulons construire vont dépendre du contexte dans lequel le mot considéré est inséré. Par exemple pour le mot *luth*, dans le contexte « jouer du », la classe qu'on cherche à construire est celle des instruments de musique mais si on s'intéresse à l'énoncé *poser un luth*, la classe construite pour *luth* correspondra plutôt à une classe générale d'objets matériels. Le sens de *poser* dans *poser un luth* est en effet celui de *poser* dans *poser un objet* plutôt que celui de *poser* dans *poser ses congés* ou dans *poser une question*. Notre objectif, à terme, est de remplacer dans notre système de désambiguïsation les noms propres ou rares par leurs classes contextuelles et de retrouver ainsi le sens des verbes.

4 Des classes de sélection distributionnelle : comment ?

4.1 Données initiales

Nous travaillons sur un corpus constitué de tous les articles du journal Le Monde sur dix ans, soit 200 millions de mots[1]. L'analyseur syntaxique Syntex (Bourigault et Fabre, 2000) est utilisé pour extraire de ce corpus une liste de mots ou syntagmes[2], structurée par des relations de dépendance syntaxique. Par exemple l'énoncé « compter sur ses amis », sera analysé par Syntex de la manière suivante :



[1] Nous remercions Benoit Habert de nous avoir autorisés à travailler à partir de son corpus
[2] Syntex permet de considérer certains syntagmes tels que *chef d'état, groupe financier* ou *Parc des Princes* comme des unités à part entière.

A partir de cette analyse, Syntex nous fournit la liste des mots lemmatisés contenus dans le corpus avec leur fréquence ainsi que la liste des *triplets* {recteur ; relation ; régi} du corpus, avec leur fréquence. On a par exemple le triplet {compter(V) ; PREP_SUR ; ami(N)}[3] dont la fréquence est 13. Il y a 20 millions de triplets différents (20 125 540 très exactement). Nous appellerons contexte lexico-syntaxique (C.L.S.) le couple formé par un des mots du triplet et la relation syntaxique. Chacun des triplets va être séparé en un C.L.S. régi, un C.L.S. régissant, et deux mots. Le triplet {compter(V) ; PREP_SUR ; ami(N)} donnera ainsi deux C.L.S., « compter(V).PREP_SUR » présent 8860 fois dans le corpus et « PREP_SUR.ami(N) » présent 88 fois et deux mots *compter(V)* et *ami(N)* de fréquences respectives 81485 et 38856. Nous obtenons ainsi une liste de mots (ou syntagmes) et une liste de contextes lexico-syntaxiques munis de leurs fréquences respectives. Ces listes constituent nos données de départ.

4.2 Données filtrées :

Les listes ainsi obtenues constituent une base de données colossale difficilement exploitable en l'état. Pour des raisons de taille et surtout de fiabilité nous avons dû filtrer les informations qu'elle contient. Nous avons appliqué successivement les critères suivants : chaque mot et chaque C.L.S. doivent être présents au moins 100 fois dans le corpus et chaque triplet doit être présent au moins 10 fois dans le corpus.

Après filtrage le corpus contient 31417 mots et 61202 contextes. A partir de ces données, nous construisons l'espace multidimensionnel engendré par les C.L.S.. C'est ce que nous appelons l'espace distributionnel associé au corpus. Chaque mot y est représenté par un point. La coordonnée d'un mot M sur l'axe engendré par un contexte C est la fréquence relative du triplet formé par M et C. Cet espace est muni de la distance du Chi2 : soit n le nombre de mots, p le nombre de contextes, M_i et $M_{k'}$ des mots de coordonnées (x_i^j) et (x_k^j) alors

$$d(M_i, M_k)^2 = \sum_{j=1}^{p} \frac{1}{x_\bullet^j} \left(\frac{x_i^j}{x_i^\bullet} - \frac{x_k^j}{x_k^\bullet} \right)^2 \text{ où } x_\bullet^j = \sum_{i=1}^{n} x_i^j \text{ et } x_i^\bullet = \sum_{j=1}^{p} x_i^j$$

4.3 Etude d'un mot dans un contexte lexico-syntaxique donné

Soient les énoncés *descendre le Mont-blanc* et *descendre la Seine*. Imaginons que nous cherchons à désambiguïser le verbe *descendre.* Une des forces de notre méthode est qu'elle permet d'étudier des cooccurrences non présentes dans le corpus. Par exemple, on ne rencontre aucune occurrence de *Mont-blanc* ni *Seine* qui soit objet de *descendre*. Il est cependant possible d'étudier les mots *Mont-blanc* et *Seine* dans le C.L.S. « descendre.OBJ ». Nous allons d'abord chercher dans l'espace distributionnel tous les mots qui ont une coordonnée non nulle selon cette dimension, c'est à dire tous les mots du corpus filtré employés dans ce contexte. On ajoute à la liste des mots obtenus les mots *Mont-blanc* et *Seine.* Notons que l'on fait une recherche toutes catégories confondues et que l'ensemble recherché peut contenir aussi bien des adjectifs, des noms communs, des noms propres ou même des entités plus complexes.

Si cet ensemble contient plus de 100 mots, on ne prend que les 100 mots les plus proches (au sens du Chi2) de *Mont-blanc* dans l'espace distributionnel. Notons MOTS l'ensemble formé. On va ensuite recenser tous les contextes pour lesquels au moins un des éléments de MOTS a

[3] Pour les relations prépositionnelles, les deux triplets {compter(V) ; __ ; sur(Prep)} et {sur(Prep) ; __ ; ami(N)} sont fusionnés en un seul {compter(V) ; PREP_SUR ; ami(N)}

une coordonnée non nulle. Notons CONT l'union de tous ces contextes. Dans le cas de *Mont-blanc*, MOTS contient 24 mots et CONT contient 5762 contextes. Nous pouvons dans un premier temps visualiser l'ensemble MOTS grâce à une analyse factorielle des correspondances (AFC) qui nous fournit 10 axes de visualisation synthétisant le mieux l'information des 5762 contextes de CONT.

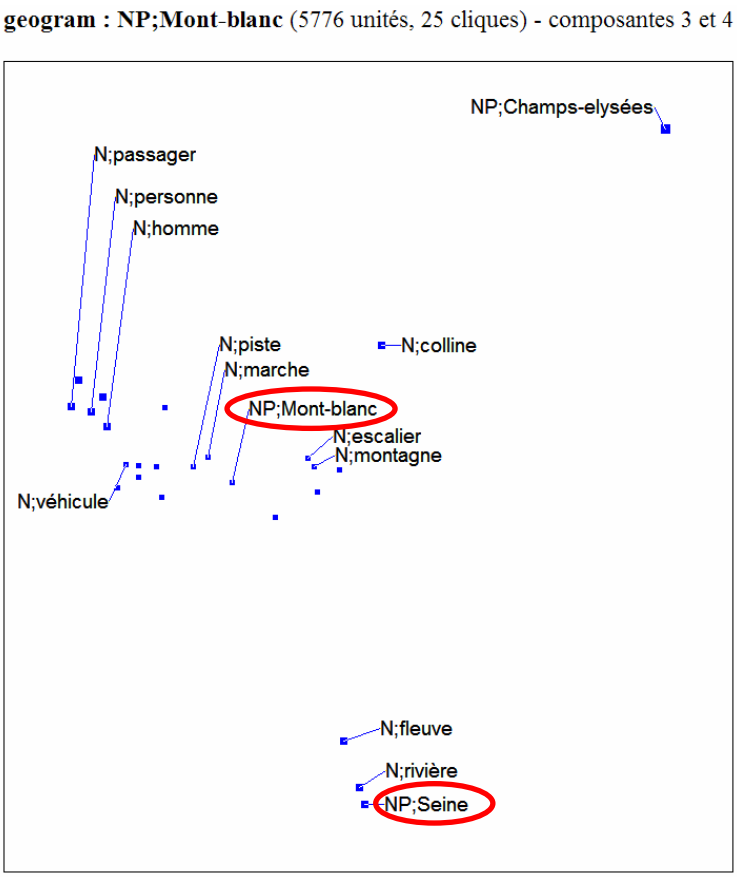


Figure 1 : distribution des mots compléments d'objet de *descendre* auxquels on a ajouté *Seine* et *Mont-blanc*

La Figure 1 fait clairement apparaître trois axes sémantiques organisant les compléments d'objet du verbe *descendre* : les personnes, les monts ou surfaces inclinées, et les cours d'eau. Notons que *descendre un avion* se trouve entre *descendre une personne* et *descendre les marches*. Il est remarquable que *Seine* et *Mont-blanc* qui ne sont pas des compléments d'objet de *descendre* dans le corpus étudié trouvent automatiquement leur place le long de l'axe qui leur correspond le mieux. *Seine* est placé à coté de *rivière* et *fleuve*, alors que *Mont-blanc* est entouré de *piste*, *montagne* et *escalier*. La visualisation proposée correspond aux composantes 3 et 4 de l'AFC. Autrement dit, l'information est contenue dans l'ensemble des composantes de l'AFC, et obtenir une visualisation lisible nécessite de parcourir les différentes composantes. C'est pourquoi la construction des classes distributionnelles va tenir compte des dix premières dimensions de l'AFC. On pondère lors de la clusterisation chaque axe par le coefficient $1-((x-1)^2/100)$ (où x est le numéro de l'axe). Nous avons choisi la clusterisation K-mean de Matlab (Seber, 1985). K-mean emploie un algorithme itératif en deux phases dont le but est de minimiser la somme des distances entre points et centre de gravité sur le nombre k de clusters. Pour un mot M et un contexte lexico-syntaxique C donnés, notre modèle propose un ensemble de classes de sélection distributionnelle employées avec le contexte C. Nous avons la possibilité d'ordonner ces classes en fonction de leur proximité avec le mot M, la première classe sémantique étant celle qui contient le mot M. Il arrive parfois que M soit l'unique mot de la classe, dans ce cas nous considérons que la classe la plus proche de M est la seconde. Ainsi dans le contexte « descendre.OBJ » la classe la plus proche de *Seine* est «N ;fleuve, N ;rivière, NP ;Seine » et la classe la plus proche de *Mont-blanc* est « N;montagne, N;piste ».

5 Evaluation des résultats

Nous proposons maintenant une première évaluation de notre système. Elle porte sur quatre contextes particulièrement ambigus en fonction des caractéristiques sémantiques de leurs

arguments : « descendre.OBJ », « jouer.PREP_à », « regarder.OBJ », « décider.SUJ ». Pour chaque contexte, nous calculons la classe sémantique la plus proche de quinze mots vedettes différents. Voici la liste des 60 cooccurences étudiées. Nous avons marqué d'un [P] celles qui sont présentes dans le corpus :

« descendre.OBJ» : NP;Seine, NP;Rhône, NP;Gange, NP;Danube, NP;Mississippi, NP;Chirac, NP;Jospin, NP;Pdg, NP;Kennedy, NP;Mont Blanc, NP ;Everest, NP;Pyrénées, NP;Alpes, NP; Broadway

« jouer.PREP_à » NP;Monopoly[P] , N;tarot, N;domino[P] , N;lego, NP;Paris[P] , NP;Washington, P;Wimbledon[P] P;Lyon, NP;Broadway, NP;New York[P], NP;Londres[P], NP;Marseille[P], NP;Lille, NP;Parc des princes[P]

« décider.SUJ » :NP;Paris[P], NP;France[P], NP;Washington[P], NP;Wimbledon, NP;Londres[P], NP;Clinton[P], NP;président, NP;Jospin[P], NP; Kennedy, NP;Onu[P], NP;Cgt[P], NP;Otan[P], NP;Rpr[P], NP ;PS[P], NP;Vivendi, NP;Renault[P]

« regarder.OBJ » : NP;Chirac, NP;Picasso, NP;Seine, NP;Arte, NP; Kennedy, NP;Internet, NP;Alpes, NP;Jospin, NP;Paris, NP;Lyon, NP;Lelouch, NP;Kubrick, NP;Etats-Unis, NP;Tf[P], NP;Tintin, N;Videocassette

L'évaluation, inspirée des travaux de Lin et Pantel (2001), consiste à juger si la classe proposée par le modèle est acceptable ou non. Huit juges, dont nous nous sommes naturellement exclus, vont donner une note de un à quatre, de la manière suivante. La classe est très mauvaise : 1 ; La classe est assez mauvaise : 2 ; La classe est assez bonne : 3 ; La classe est très bonne : 4.

Contexte « descendre.OBJ »		
Mots vedettes	**Cluster le plus proche du mot vedette**	**Ev.**
NP;Seine	N;fleuve, N;rivière, N;Seine	4
NP;Gange	N;fleuve, N;rivière	3,9
NP;Chirac	N;homme, N;personne	3,8
NP;Pdg	N;homme, N;personne	3,1
NP;Mont Blanc	N;montagne, N;piste	3,5
NP ;Pyrénées	N;rivière, N;fleuve	1,6
Contexte « jouer.PREP_à »		
Mots vedettes	**Cluster le plus proche du mot vedette**	**Ev.**
NP;Monopoly	N;billard, N;bridge, N;cache-cache, N;domino, N;ping-pong, N;poker, N;pétanque, N;souris, N;yo-yo, NP;Monopoly	3,4
NP;Lego	N;billard, N;bridge, N;cache-cache, N;domino, NP;Lego, N;ping-pong, N;poker, N;pétanque, N;yo-yo, NP;Monopoly	3,5
NP;Paris	NP ;New York, NP;Avignon, NP;Londres, NP;Marseille, NP;Paris	3,5
NP;Washington	NP ;New York, NP;Londres, NP;Paris, NP;Washington	3,6
NP;Wimbledon	N;basket, N;basket-ball, N;football, N;loterie, N;rugby, N;tennis, S;jeu vidéo	1,9
NP;Broadway	N;base-ball, N;cricket, N;foot, N;golf, N;loto, N;volley-ball, NP;Broadway	1,4
Contexte « décider.SUJ»		
Mots vedettes	**Cluster le plus proche du mot vedette**	**Ev.**
NP;Washington	NP;Etats-Unis, NP;Washington	3.7
NP;Wimbledon	NP;Europe, NP;France, NP;Italie	2
NP;président	NP;Pdg, S;directeur général, S;président de le conseil, S;secrétaire de état	3.7
NP;Chirac	NP;Chirac, NP;Clinton, NP;Eltsine, NP;Jospin, NP;Mitterrand, S;chef de le état, S;premier ministre, S;président de le république	3.4
NP;Otan	N;armée, N;force, N;police	3.4
Contexte « regarder.OBJ »		
Mot vedettes	**Cluster le plus proche du mot vedette**	**Ev.**
NP;Chirac	N;enfant, N;femme, N;gens, N;homme, N;personne, N;public, NP;Chirac	3.1
NP;Seine	N;mer	2.8
NP;Arte	NP;Arte, NP;Tf, S;chaîne de télévision	3.8
NP;Alpes	N;montagne	3.8

Figure 2 : mots évalués dans le contexte « descendre.OBJ »

Nous avons détaillé quelques notes pour chaque contexte dans la Figure 2. Dans le contexte « descendre.OBJ », on peut constater que les noms propres *Gange* et *Seine* ont dans leur première classe *fleuve* et *rivière*. Les noms propres *Chirac* et *PDG* ont dans leur première classe *homme* et *personne*. Dans le contexte « jouer.PREP_à », des noms propres comme *Lego* ou *Monopoly* sont rattachés à des classes de jeux alors que *Paris* et *Washington* sont rattachés à des noms de villes.

On peut noter des erreurs telles que « descendre.OBJ » avec *Pyrénées* qui donne des classes contenant *fleuve* et *rivière*. Ou encore « jouer.PREP_à » avec *Broadway*, qui donne une classe contenant des noms de sport.

	jouer.à	descendre.Obj	regarder.OBJ	décider.Suj	**TOTAL**
% 3 et 4	**81,63**	**80,61**	**65.18**	**83.93**	**77.62**
4	48,98	57,14	31.25	63.39	**50.00**
3	32,65	23,47	33.93	20.54	**27.62**
2	5,10	10,20	25	13.39	**13.81**
1	8,16	8,16	9.82	2.68	**7.14**
Non réponse	5,10	1,02	0	0	**1.43**
Moyenne des notes	3.3	3.2	2.9	3.4	**3.2**
Moyennes >3 (en %)	85.71	57.14	31.25	68.75	**60**

Figure 3 : Evaluation des classes obtenues pour 4 contextes, 60 cooccurrences, 8 juges.

La Figure 3 propose une synthèse des résultats sur les quatre contextes. Les résultats de l'évaluation sont tout à fait satisfaisants. 77.62 % des notes sont supérieures ou égales à 3 et correspondent donc à des jugements de classe assez ou très bonnes. On peut noter que dans la moitié des cas la note mise est un 4. La moyenne des notes sur les 4 contextes est de 3,2 ce qui veut dire que les classes proposées sont globalement attestées par les juges. Enfin 60% des classes ont sur l'ensemble des juges une note moyenne strictement supérieur à 3. On peut donc dire que 60% des classes proposées par notre système sont bonnes ou assez bonnes. 5 d'entre elles ont été jugées très bonnes par tous les juges. Ce sont les classes correspondant aux énoncés « descendre la Seine/ le Rhône/ le Danube » et « le PS/ la CGT décide ».

Il est par ailleurs intéressant d'étudier le comportement d'un même nom propre dans des contextes différents. Par exemple, *Wimbledon* peut être employé dans des énoncés tels que *revenir de Wimbledon, Wimbledon décide* ou *jouer à Wimbledon*. La Figure 4 présente pour chacun des contextes correspondants, les deux classes les plus proches de *Wimbledon*. On observe que les différentes facettes de *Wimbledon* sont mises en évidence en fonction du contexte. Le contexte «décide.SUJ » met en valeur *Wimbledon* en tant que zone géographique de décision. Le contexte « jouer.PREP_à» insiste sur la singularité de *Wimbledon* en tant que compétition de tennis. Les classes obtenues pour le contexte « revenir.PREP_de » peuvent être interprétées comme des lieux d'activité.

	Wimbledon dans le contexte « Wimbledon décide»	**Wimbledon** dans le contexte « jouer à Wimbledon »	**Wimbledon** dans le contexte « revenir de Wimbledon »
1er cluster	NP;Europe, NP;France, NP;Italie	N;basket, N;basket-ball, N;football, N;loterie, N;rugby, N;tennis, S;jeu vidéo	NP;Allemagne, NP;Etats-unis
2ème cluster	N;monde, N;pays, N;région, N;ville	N;base-ball, N;cricket, N;foot, N;golf, N;loto, N;volley-ball	N;guerre, N;mission, N;travail

Figure 4 : différentes classes de *Wimbledon* en fonction du contexte

6 Conclusions et perspectives

Nous avons présenté ici une méthode de construction de classes de sélection distributionnelle en contexte. Cette méthode, au vu des résultats préliminaires présentés ici, nous semble très prometteuse. Des expérimentations plus poussées sont cependant encore nécessaires pour finaliser notamment la méthode de clusterisation (détermination du nombre de clusters optimal, pondération des axes de l'AFC) ou le mode de filtrage du corpus. Il nous faudrait valider sur un plus grand nombre de contextes, sur différentes catégories de mots et en faisant appel à un plus grand nombre de juges. Nous devons aussi étudier la variation des classes obtenues lorsqu'on change le corpus de travail. Le type d'évaluation que nous avons commencé à mettre en place n'est qu'une étape. Le but à terme est d'utiliser les classes obtenues, dans notre système de désambiguïsation. Il nous faut donc mettre au point le module de Visusyn correspondant. Nous devrons déterminer quelle est la façon la plus pertinente de prendre en compte les classes de sélection distributionnelle dans Visusyn. On peut envisager d'injecter l'ensemble d'une classe, de lui associer une région dans l'espace sémantique du mot à désambiguïser, mais on peut aussi imaginer de travailler avec un représentant de la classe, une sorte de prototype. Se pose alors la question du statut du prototype : est ce le mot le plus proche du centre de gravité ou bien le centre de gravité lui-même ? Dans le dernier cas, le prototype ne serait pas un item lexical mais une sorte d'abstraction sans dénomination linguistique qu'on ne pourrait appréhender que par son profil d'utilisation (il se construit à 26% avec descendre.OBJ, à 10% avec revenir.DE, …) Ce sont les résultats de ce module, après évaluation de leur fiabilité, qui constitueront la véritable validation des classes obtenues. Les jugements des locuteurs humains se réfèrent en effet à une utilisation quotidienne du français. Ainsi certaines classes qui nous semblaient pertinentes dans une tâche de désambiguïsation sont rejetées massivement. C'est le cas par exemple de la classe formée pour *Wimbledon* dans le contexte « jouer.PREP_à ». Il nous semblait que le fait d'obtenir des noms de sport était satisfaisant puisque cela permet de donner à *jouer* le sens de « *pratiquer un sport* ». Or la moyenne des notes de cette classe n'est que de 1,9.

Une autre perspective de travail est de quitter le champ de la désambiguïsation pour celui de la catégorisation. Un de nos résultats importants est de montrer que la classe d'un même mot varie avec le contexte. Le fait par exemple que *Wimbledon* soit catégorisé parfois comme une zone géographique de décision et d'autre fois assimilé à un sport nous donne à penser que la langue n'est pas organisée selon un système hiérarchique de classes fixes, même si on accepte les recouvrements de classes. La catégorisation d'un mot se ferait plutôt « à la volée », de façon dynamique et en contexte.

Remerciements

Nous remercions les personnes qui ont accepté d'évaluer nos résultats, Sophie Prévost, Laure Sarda et Bernard Victorri pour leurs commentaires avisés, Benoît Habert pour la mise à disposition de son corpus, et plus particulièrement Didier Bourigault qui nous a fourni nos données de départ ainsi que de précieux conseils.

Références

AUSSENAC-GILLES N., BIÉBOW B., SZULMAN N. (2000), Revisiting Ontology Design: a method based on corpus analysis, Actes de 12th International Conference on Knowledge Engineering and Knowledge Management. Juan-Les-Pins

BOURIGAULT D. (2002), Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, Actes de *TALN 2002*, Nancy, pp. 75-84

BOURIGAULT D., FABRE C. (2000), Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaires,* n° 25, Université Toulouse - Le Mirail, pp. 131-151.

FRANCOIS F., MANGUIN J.L., VICTORRI B. (2003), *La réduction de la polysémie adjectivale en cotexte nominal : une méthode de sémantique calculatoire*, Cahier du Crisco n°14

FLEURY S. (1998), Gaspar, un dispositif de TALN basé sur la programmation à Prototypes, Actes de TALN'98, Paris.

FREROT C., BOURIGAULT D, FABRE C. (2003), Marier procédures d'apprentissage endogènes et ressources exogènes dans un analyseur syntaxique de corpus – Le cas du rattachement verbal à distance de la préposition « de » , *T.A.L*, 44-3.

GOLDBERG A. (1995), *Constructions : a construction grammar approach to argument structure*, Chicago and London, University of Chicago Press.

GREFENSTETTE G (1994)., *Explorations in Automatic Thesaurus Discovery*, London, Kluwer Academic Publishers.

GROSS G (2004), Réflexions sur le traitement automatique des langues, Actes de *JADT 2004,* Vol. 1 545-556 .

HABERT B., ILLOUZ G., FOLCH H. (2004), Dégrouper les sens: pourquoi, comment?, Actes de *JADT 2004*, Vol. 1 565-576 .

HABERT B., FOLCH H. ET ILLOUZ G. (1999). Sortir des sens uniques : repérer les mots « mouvants » dans le domaine social. *Sémiotiques*, vol. (17). *Dépasser les sens iniques dans l'accès automatisé aux textes*, Habert B. (resp.) : 121-151.

HABERT B., NAZARENKO A. (1996)., La syntaxe comme marche-pied de l'acquisition des connaissances : bilan critique d'une expérience, *Journées sur l'acquisition des connaissances*, AFIA, Sète

JACQUET G. (2004), Using the construction grammar model to disambiguate polysemic verbs in French, Actes de *ICCG3 (International Conference on Construction Grammar),* Marseille.

JACQUET G. (à paraître), A model of disambiguation of polysemic verbs in French, *Constuctions, http://www.constructions-online.de/*

KAY P. (2000), Argument Structure Constructions and the Argument-Adjunct Distinction, Actes de *ICCG1*, Berkeley, p 30.

LIN D., PANTEL P. (2001), Induction of Semantic Classes from Natural Language Text, Actes de *ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001*.

PLOUX S., VICTORRI B. (1998), Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes, *Traitement automatique des langues*, Vol. 39, n°1, pp.161-182.

SEBER G.A.F. (1984), *Multivariate Observations*, Wiley, New York. pp. 317-322.

VENANT F. (2004), Polysémie et calcul du sens, Actes de *JADT 2004,* Vol 2. 1146-1157.

VICTORRI B., FUCHS C. (1996), *La polysémie, construction dynamique du sens*, Paris, Hermès.