Utiliser des classes de sélection distributionnelle pour désambiguïser les adjectifs

Fabienne VENANT
LaLIC – Université Paris IV, Maison de la recherche,
28 rue Serpente 75006 Paris
fabienne.venant@ens.fr

Résumé. La désambiguïsation lexicale présente un intérêt considérable pour un nombre important d'applications, en traitement automatique des langues comme en recherche d'information. Nous proposons un modèle d'un genre nouveau, fondé sur la théorie de la construction dynamique du sens (Victorri et Fuchs, 1996). Ce modèle donne une place centrale à la polysémie et propose une représentation géométrique du sens. Nous présentons ici une application de ce modèle à la désambiguïsation automatique des adjectifs. La méthode utilisée s'appuie sur une pré-désambiguïsation du nom régissant l'adjectif, par le biais de classes de sélection distributionnelle. Elle permet aussi de prendre en compte les positions relatives du nom et de l'adjectif (postpostion ou antéposition) dans le calcul du sens.

Abstract. Automatic word sense disambiguation represents an important issue for many applications, in Natural Language Processing as in Information Retrieval. We propose a new kind of model, within the framework of Dynamical Construction of Meaning (Victorri and Fuchs, 1996). This model gives a central place to polysemy and proposes a geometric representation of meaning. We present here an application of this model to adjective sense disambiguation. The method we used relies on a pre-disambiguation of the noun used with the adjective under study, using distributionnal classes. It can also take into account the changes in the meaning of the adjective, whether it is placed before or after the noun.

Mots-clés : traitement automatique des langues, désambiguïsation, sémantique, polysémie adjectivale, construction dynamique du sens, synonymie, classes distributionnelles, corpus, espace sémantique, espace distributionnel.

Keywords : natural language processing, word sense disambiguation, semantics, adjectival polysemy, dynamical construction of meaning, synonymy, distributionnal classes, corpus, semantic space, distributional space.

1 Enjeux actuels pour la désambiguïsation automatique

Le développement et la généralisation de l'utilisation de documents numériques génèrent des nouveaux besoins en analyse textuelle, notamment pour la navigation dans des bases de données numériques et la recherche d'information. La désambiguïsation automatique constitue

une étape importante dans l'analyse de ces données textuelles, dans les phases d'indexation, ou de description, des documents, comme dans celle de l'extension de requête. Un des enjeux dans ce domaine est, en effet, de pouvoir faire des requêtes non pas sur des occurrences de mots simples mais sur des concepts, et donc de construire des outils capables de prendre en compte la polysémie. Le but du travail présenté ici est donc d'évaluer les performances d'un modèle dynamique de calcul du sens, et des outils géométriques qu'il met en jeu, dans des tâches de désambiguïsation sémantique d'assez haut niveau. La polysémie adjectivale constitue pour cela un champ d'expérimentation idéal, assez peu exploré d'un point de vue informatique. Les travaux existants portent surtout sur la catégorisation des adjectifs, en lien avec l'acquisition lexicale ((Rasking et Nurinburg, 1996) ; (Bouillon et Viegas, 1999)). Les travaux en désambiguïsation automatique se sont quant à eux majoritairement intéressés aux noms et aux verbes. Les phénomènes mis en jeu dans la sémantique adjectivale sont en effet très subtils, difficiles à formaliser et à expliquer de façon systématique. La prise en compte de ces phénomènes peut cependant beaucoup apporter à des outils de recherche d'information, notamment dans des études d'opinions ou de modalités sentimentales.

2 Un modèle dynamique du sens

Le système informatique que nous avons développé met en jeu le modèle dynamique du sens proposé par Victorri et Fuchs (1996). Le principe est le suivant : on associe à chaque mot polysémique un espace sémantique dans lequel se déploient ses différents sens. Les autres mots présents dans l'énoncé définissent une fonction potentielle, et ce sont les sommets de cette fonction potentielle qui permettent de déterminer le sens pris par le mot étudié, dans l'énoncé considéré. Le développement du système s'est fait en deux étapes. Un premier travail, portant sur la représentation du sens, a consisté à mettre au point une méthode de construction automatique des espaces sémantiques et d'exploration du lexique, par le biais de la relation de synonymie ((Ploux et Victorri, 1998) ; (Venant 2007)). Nous présentons cette méthode en section 3. Un second travail a permis d'utiliser ces espaces sémantiques dans une méthode de désambiguïsation automatique. Les travaux ont porté d'une part sur la prise en compte de la polysémie verbale (Jacquet, 2006), d'autre part sur la prise en compte de la polysémie adjectivale (Venant, 2006).

Nous présentons ici les dernières avancées dans la désambiguïsation adjectivale. Les études linguistiques sur l'adjectif s'accordent pour dégager deux caractéristiques principales de la sémantique adjectivale. La première concerne le fait que le sens de l'adjectif dépend du nom qui le régit. Ainsi *sec* prend des sens différents dans *un terrain sec* et *un visage sec.* La seconde concerne l'influence sur le sens de l'adjectif de sa position relativement au nom. Ainsi *un curieux homme* n'est pas nécessairement *un homme curieux*. Nous attendons du système qu'il soit capable de prendre en compte ces caractéristiques dans le calcul du sens d'un adjectif en présence d'un nom donné. Avant de lancer une étude à grande échelle, nous avons voulu étudier la plausibilité du système par une étude en profondeur des adjectifs *sec*, *curieux* et *méchant*. Notre travail s'appuie une pré-désambiguïsation du nom par le biais de classes de sélection distributionnelle. La section 4 présente la méthode de construction automatique de ces classes. Les sections 5 et 6 détaillent la façon dont nous utilisons ces classes, ainsi que les espaces sémantiques, pour prendre en compte l'influence du nom recteur dans le calcul du sens d'un adjectif. La section 7 porte sur le traitement des changements de sens entre anté et postposition.

3 Construction des espaces sémantiques

Nous illustrons ici, sur le cas de l'adjectif *méchant*, la méthode de construction des espaces sémantiques mise au point par Ploux et Victorri (1998). Cette méthode repose sur l'analyse d'un graphe de synonymie (Dictionnaire Electronique des Synonymes, DES : www.crisco.unicaen.fr). Le DES fournit le graphe de synonymie de *méchant* : les sommets du graphe sont *méchant* et tous ses synonymes, et il y a un lien entre deux de ces adjectifs lorsque le DES indique un renvoi synonymique. La Figure**Erreur ! Aucun nom n'a été donné au signet.Erreur ! Aucun nom n'a été donné au signet.** montre un extrait du graphe de synonymie de *méchant*.

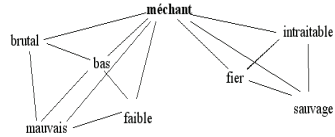


Figure 1 : un extrait du graphe de synonymie de *méchant*

Un synonyme ne suffit pas en général pour définir un sens du mot étudié. On voit, par exemple, sur la Figure **Erreur ! Aucun nom n'a été donné au signet.Erreur ! Aucun nom n'a été donné au signet.** que *bas* est à la fois synonyme de *brutal* et de *faible*, qui correspondent à deux sens différents de *méchant*. On va donc utiliser des ensembles de synonymes, et plus précisément les cliques du graphe. Une clique est un ensemble de sommets deux à deux synonymes le plus grand possible. Le graphe de la Figure **Erreur ! Aucun nom n'a été donné au signet.** présente ainsi 3 cliques : *<bas ; brutal ; mauvais ; méchant>, <bas ; faible ; mauvais* ; *méchant>* et *<fier ; intraitable ; méchant ; sauvage>*. On va considérer qu'une clique correspond, en première approximation, à une nuance de sens possible pour le mot considéré. Ce sont donc les cliques qui constitueront les points de l'espace sémantique. L'espace sémantique est alors défini comme l'espace euclidien engendré par *méchant* et tous ses synonymes. A chaque adjectif correspond un axe de l'espace. A chaque clique du graphe correspond un point de l'espace, dont les coordonnées dépendent des synonymes qu'elle contient. Cet espace est muni de la distance du Chi2, bien connue en analyse des données, de façon à rendre compte des proximités sémantiques réelles entre les différents sens du mot étudié. On utilise une Analyse en Composantes Principales pour obtenir une visualisation en deux ou trois dimensions. La Figure **Erreur ! Aucun nom n'a été donné au signet.** présente la visualisation de l'espace sémantique associé à *méchant*. L'espace construit automatiquement rend correctement compte de la sémantique de l'adjectif *méchant*. En effet, on peut constater que les sens de *méchant* se répartissent en trois zones, correspondant aux distinctions de sens apparaissant dans les dictionnaires. En haut à gauche, on trouve les sens intensifs (intensivité négative: *incapable*, *dérisoire, déficient...*), les plus généraux. La partie droite de l'espace sémantique organise les sens les plus spécifiques de *méchant*. En haut à droite, on trouve les cliques correspondant aux sens s'appliquant surtout à des personnes et à leurs actes. En bas à droite on trouve regroupées les cliques correspondant à des sens psychologiques de *méchant*, caractérisant par exemple des attitudes ou des sentiments.

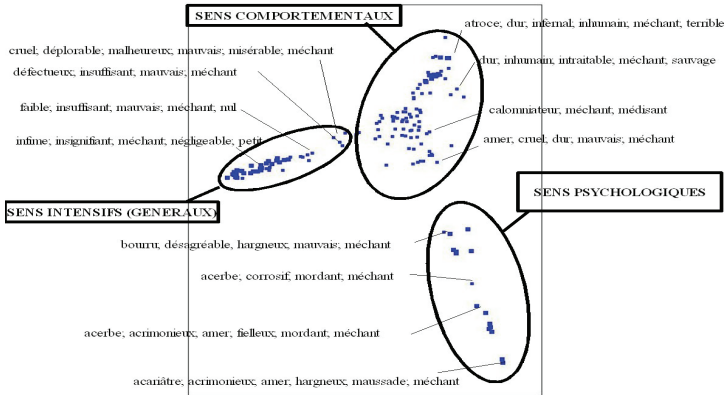


Figure 2 : Espace sémantique associé à *méchant*

4 Utiliser des classes distributionnelles

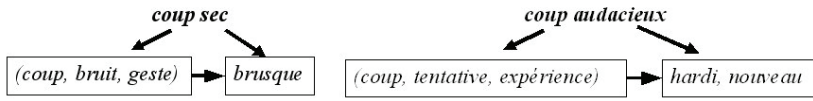


Figure 3: Influence mutuelle nom-adjectif

Nous avons, dans un travail antérieur (Jacquet et Venant, 2005), mis au point une méthode de construction automatique de classes de sélection distributionnelle (CSD), à partir d'un corpus. Le principe de construction des CSD est de rassembler dans une même classe les noms influençant de la même manière le sens d'un adjectif. Les noms sont rassemblés sur la base des contextes lexico-syntaxiques qu'ils partagent. Un contexte lexico-syntaxique est constitué d'un mot et d'une relation syntaxique, comme par exemple être recteur de l'adjectif *sec* en position épithète (codé *sec*.EPI), ou être complément d'objet direct du verbe donner (codé *donner*.OBJ). Ces classes vont servir ici à rendre compte de l'aspect dynamique du calcul du sens, en permettant une double désambiguïsation nom-adjectif. Les CSD constituent en effet un outil adéquat pour la prise en charge de l'influence de l'adjectif sur la sémantique du nom qui le régit. Considérons ainsi le groupe nominal *coup sec*. On va ainsi définir une classe (*coup, bruit, geste*..), rassemblant des noms en présence desquels *sec* prend un sens dénotant un manque de douceur. Cette classe se distingue d'une part de celles d'autres noms, comme par exemple celle associée au nom *fruit,* rassemblant des noms (*fruit, haricot, légume*...), en présence desquels *sec* prend un sens dénotant un manque d'eau. Elle se distingue d'autre part de celle associée au même nom *coup*, dans l'étude d'un autre syntagme, par exemple le syntagme *coup audacieux* (*cf.* figure 3). Ce qui nous intéresse particulièrement ici, c'est que la classe d'un nom varie en fonction de l'adjectif étudié, et que c'est cette classe qui permet

ensuite de désambiguïser l'adjectif, et de lui assigner le sens correct en présence du nom considéré.

Pour construire automatiquement ces classes, nous travaillons à partir des sorties de l'analyseur syntaxique Syntex (Bourigault et Fabre, 2000) sur le corpus *Le Monde LM10*. Après filtrage, le corpus contient 31 417 mots et 61 202 contextes lexico-syntaxiques (CLS). A partir de ces données, nous construisons l'espace multidimensionnel engendré par les CLS. C'est ce que nous appelons l'espace distributionnel associé au corpus. Chaque mot y est représenté par un point. Les coordonnées de ce point dépendent des fréquences d'emploi du mot associé dans chacun des CLS engendrant l'espace. La classe d'un nom N en tant que recteur d'un adjectif A rassemble des noms qui sont eux aussi attestés comme étant recteur de l'adjectif A, et qui sont proches du nom N dans l'espace distributionnel. Nous ne détaillons pas ici la méthode de construction de la classe (*cf.* Venant 2006), mais nous l'illustrons sur le cas du nom *bruit.* Pour construire la classe associée au nom *bruit* en tant que recteur de *sec* dans une relation épithète, on commence par chercher dans l'espace distributionnel tous les noms attestés comme recteur de *sec* dans le corpus, c'est-à-dire les noms qui ont une coordonnée non nulle selon la dimension *sec*.EPI. Si cet ensemble contient plus de 100 mots, on ne prend que les 100 mots les plus proches (au sens du Chi2) de *bruit* dans l'espace distributionnel. Notons MOTS l'ensemble formé. On va ensuite recenser tous les contextes pour lesquels au moins un des éléments de MOTS a une coordonnée non nulle, c'est-à-dire l'ensemble des contextes dans lesquels au moins un des noms inclus dans MOTS est employé. Notons CONT l'union de tous ces contextes. Dans le cas de *bruit*, MOTS contient 59 mots (l'ensemble des noms recteurs de *sec* dans le corpus) et CONT contient 9 506 contextes. Une Analyse en Composantes Principales fournit alors les 10 axes de visualisation synthétisant le mieux l'information des 9 506 contextes de CONT, ainsi que les coordonnées des points représentant les 59 mots étudiés dans l'espace euclidien engendré par ces 10 axes. On peut ainsi obtenir des représentations en deux ou trois dimensions de l'espace engendré par CONT. La figure 4, ci- dessous, montre ainsi deux visualisations de l'ensemble des noms recteurs de *sec* dans l'espace distributionnel. On peut ainsi s'apercevoir que *bruit* est proche de *coup* et *geste*, et que *fruit* est proche de *légume* et *pain.* Chaque nom est ainsi proche de noms en présence desquels l'adjectif *sec* prend des sens similaires.

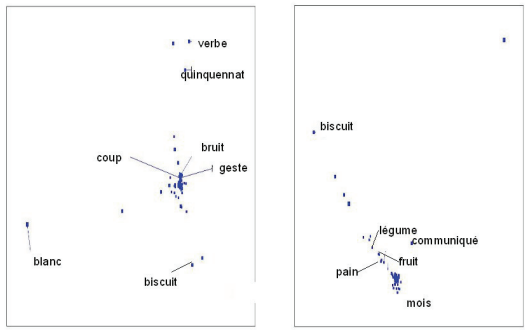


Figure 4: Noms recteur de *sec* dans l'espace distributionnel

Pour constituer la classe distributionnelle de *bruit*, dans le contexte *sec*.EPI, le système classe les noms selon leur distance au nom étudié. Il ajoute ces noms à la classe, du plus proche au

plus éloigné, en additionnant leurs fréquences respectives dans le corpus. Il continue tant qu'un certain seuil de présence dans le corpus n'a pas été atteint. Ce seuil correspond au troisième quartile de la série des fréquences dans le corpus des noms étudiés. Il vaut ici 10 960. La classe distributionnelle de *bruit*, en tant que recteur de l'adjectif épithète *sec*, est (*bruit,* coup).

5 Influence du nom recteur

Nous avons, dans un premier temps, travaillé sur l'influence du nom recteur sur le sens de l'adjectif *sec*. L'influence du nom est prise en compte sous la forme d'une fonction potentielle, définie sur l'espace sémantique de l'adjectif étudié, et dont les sommets permettent de déterminer dans quelle zone de l'espace sémantique ce nom contraint l'adjectif à prendre son sens. Il s'agit donc ici d'associer à chaque nom, ou plutôt à la CSD de ce nom en tant que recteur de *sec*, un potentiel désambiguïsateur. Pour cela, le système que nous avons développé calcule le degré d'affinité de la CSD avec chacune des cliques de *sec,* en fonction des fréquences de cooccurences de chacun des noms présents dans la classe avec chacun des synonymes présents dans la clique. Le tableau 1 présente quelques degrés d'affinités, entre certaines cliques de *sec* et la classe C : (*bruit; coup*).

CLIQUES	Degré d'affinité avec C
austère ; rude ; sec ; simple	98%
bourru ; dur ; rude ; rébarbatif ; sec ; sévère	98%
bourru ; brutal ; dur ; rude ; sec ; sévère ; âpre	97%
sec ; seul ; simple	96%
bourru ; brutal ; cru ; dur ; rude ; sec	95%
bourru ; brusque ; désagréable ; rude ; sec	95%

Tableau 1. Cliques présentant les plus fortes affinités avec la classe C : (bruit, coup)

Ces degrés d'affinité sont ensuite mis en jeu dans le calcul d'une fonction potentielle associée à la classe. La valeur de la fonction en chaque point dépend du degré d'affinité de la classe avec la clique associée à ce point. La fonction potentielle associée à la classe C : (*bruit,coup*) est la suivante : elle atteint son maximum dans la région de l'espace sémantique de *sec* qui rassemble les cliques exprimant le manque de douceur (*ie.* contenant des adjectifs comme *brusque, bref, brutal*...), ce qui correspond bien au sens pris par *sec* dans le syntagme *un bruit sec.*

6 Calcul du sens d'un adjectif en présence d'un nom donné

Nous avons partitionné, manuellement, l'espace sémantique de *sec* en zones correspondant aux sens principaux distingués par les dictionnaires : le manque d'eau et l'improductivité (*une fleur sèche, un terrain sec*), le manque de douceur (*un coup sec*), la maigreur (*un visage sec)*, les sens psychologiques (*un coeur sec*)... La tâche du système consiste ensuite à déterminer automatiquement quelle est la zone de l'espace sémantique correspondant au sens pris par *sec* en présence d'un nom donné. Le système associe une fonction potentielle à chacune de ces zones. Cette fonction dépend des cliques appartenant à la zone. Le système effectue alors un calcul intégral pour comparer ces fonctions aux fonctions potentielles des noms, et déterminer comment se répartit le potentiel désambiguïsateur de chacun des noms étudiés, relativement aux différentes zones de sens. Le tableau 2 donne un aperçu partiel des résultats. La tâche a été menée sur 49 noms. Pour 26 d'entre eux, le système sélectionne la zone de sens adéquate,

pour 16 d'entre eux, le résutat est faux, c'est à dire que la zone sélectionnée n'est pas pertinente. Enfin pour les 7 noms restants, le système reste silencieux et ne sélectionne aucune zone de sens. Il s'agit d'une part des mots *mois, refus, régime* et *vol,* très fréquents dans le corpus, donc seuls dans leur classe distributionnelle*,* mais par ailleurs peu employés avec *sec* et ses synonymes, de sorte que le calcul n'aboutit pas*.* Pour les autre noms, *bois, geste* et *licenciement,* l'apport de la classe reste insuffisant en termes de fréquences de cooccurrence avec les synonymes de *sec*. Pour ces 7 noms, un deuxième calcul a été mené, après inclusion dans la classe distributionnelle du nom le plus proche dans l'espace distributionnel. Le système a pu cette fois sélectionner la zone de sens correcte, à l'exception du nom *mois* pour lequel il sélectionne la zone des sens psychologiques.

NOM	***ZONE DE SENS***	***TAUX D'AFFINITE***
sol	Manque d'eau, improductivité Manque de douceur Maigre, décharné	28% 27% 19%
humour	Sens psychologiques Manque de douceur Manque d'eau, improductivité	25% 22% 18%
cheveu	Manque d'eau, improductivité	77%
ton	Manque de douceur Sens psychologiques	42% 33%
hiver	Sens psychologiques	100%

Tableau 2 : Influence du nom recteur, quelques résultats sur l'adjectif *sec*

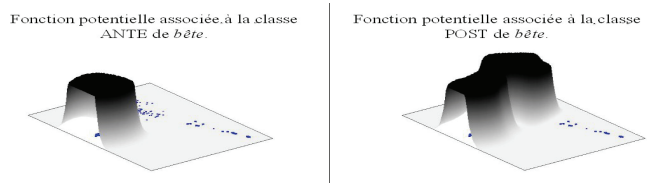
Afin de mesurer l'apport des classes distributionnelles, et donc de la désambiguïsation simultanée du nom et de l'adjectif, ces résultats ont été comparés à ceux obtenus, pour les même noms et sur la même tâche, sans utiliser les classes distributionnelles, c'est-à-dire en associant une fonction potentielle au nom seul, selon ses fréquences d'emplois avec les différents adjectifs.

Il s'avère, et c'est heureux, que l'utilisation des classes distributionnelles ne nous fait perdre aucun des résultats positifs qui apparaissent avec le nom seul. Elle apporte au contraire quelques nuances de sens, puisqu'on voit apparaître le fait qu'un *humour sec*, sec bien sûr d'un point de vue psychologique, est aussi dénué de douceur. De même, les classes distributionnelles permettent au système de détecter que le manque d'eau d'un sol en fait un sol rude, difficile à exploiter, ou encore que *sec* dans *ton sec* est porteur à la fois d'un sens acoustique et d'une connotation psychologique*.* L'utilisation des classes distributionnelles permet aussi d'obtenir des résultats corrects là où l'utilisation du nom seul génère une erreur (c'est le cas pour le nom *arbre*), ou ne donne pas de résultat du tout (14 noms sont concernés dont *bruit* et *cheveu*). Le point important est que l'utilisation des classes distributionnelles ne génère pas d'erreur supplémentaire. Les erreurs obtenues sur des noms comme *hiver* relèvent du mode de description du sens que nous utilisons, et non de l'utilisation des classes distributionnelles. Les points des espaces sémantiques sont en effet des cliques de synonymes. On se heurte ici au fait que non seulement la synonymie entre *froid* et *sec* n'est pas pertinente dans le contexte de *hiver,* mais qu'en plus cela concerne des cliques entières, puisque le phénomène se reproduit pour *glacé* et *glacial*, qui partagent de nombreuses cliques avec *froid* et *sec*. Il faut ici chercher un moyen d'informer notre système que *sec* déploie ses sens dans deux directions sémantiques, l'une plutôt physique, l'autre plutôt psychologique, mais que employé avec certains noms, comme *hiver,* il ne peut prendre son sens que dans le domaine physique, et que donc seules les cliques correspondantes sont à prendre en compte dans le

calcul du sens en présence de ces noms. Une piste pour la résolution de ce problème est l'utilisation d'informations globales sur le lexique adjectival, que nous pensons obtenir automatiquement grâce à notre travail sur la caractérisation des emplois adjectivaux (Venant, 2007).

7 Influence de la position de l'adjectif

Une première étude menée sur l'adjectif *curieux* (François, Victorri et Manguin, 2002) avait montré que le système était capable de rendre compte des changements de sens entre antéposition et postposition. Nous avons poursuivi l'investigation de ce phénomène, par l'étude de l'adjectif *méchant,* dont le sémantisme en antéposition est plus complexe que celui de *curieux*. *Méchant* possède en effet une très grande extension (il peut s'appliquer à n'importe quoi, de la table au costume en passant par l'avocat, la fée, ou la pendule). En antéposition, dans ses emplois généraux, il est sujet au phénomène de désémantisation décrit par Goes (1999), c'est-à-dire qu'il prend un sens si général que le sens du syntagme semble dépendre essentiellement du nom recteur. Enfin, les changements de sens entre antéposition et postposition ne sont pas systématiques. On trouve en effet des cas où le changement de sens est flagrant, et d'autres où le sens de l'adjectif est le même qu'il soit placé avant ou après le nom. Ainsi, s'il est préférable de ne pas engager *un méchant avocat* pour se défendre à un procès, *un avocat méchant* peut au contraire se montrer redoutable. En revanche, on redoutera de la même façon de recevoir *un coup méchant* ou *un méchant coup*. Nous attendons de notre système qu'il soit capable de rendre compte de tous ces phénomènes.
La méthode de désambiguïsation repose sur les mêmes principes que précédemment, mais on utilise ici deux classes distributionnelles [ANTE *vs.* POST] pour chaque mot, l'une calculée à partir des fréquences en antéposition, l'autre calculée à partir des fréquences en postposition. On mesure ensuite l'influence du nom recteur sur la sémantique de *méchant* en associant à chacune de ces classes une fonction potentielle, définie sur l'espace sémantique. Rappelons qu'à chaque point de l'espace sémantique correspond une clique du graphe de synonymie de *méchant*. La valeur de la fonction en chaque point dépend des fréquences de cooccurrence (en antéposition pour une classe [ANTE], en postposition pour une classe [POST]) de chacun des noms constituant la classe avec chacun des adjectifs constituant la clique (pour le détail des calculs voir Venant, 2006). La Figure 5 montre les fonctions potentielles associées aux CSD du nom *bête*.



.Figure 5 : Fonctions potentielles associées aux classes distributionnelles de *bête*

Le système doit alors déterminer le sens que prend l'adjectif *méchant* en présence d'un nom donné. Par sens, nous entendons ici, comme dans le paragraphe précédent, la zone (sens intensifs, sens comportementaux ou sens psychologiques) de l'espace sémantique correspondant au sens de *méchant* dans le syntagme étudié. Ces zones ont été définies manuellement. Le système calcule une fonction potentielle pour chacune des zones. Rappelons que, en chaque point, la valeur de la fonction dépend de l'appartenance ou non de

la clique à la zone considérée. Etant donné un groupe nominal *méchant* + N ou N + *méchant*, le système détermine la CSD [ANTE] ou la CSD [POST], associée à N en tant que recteur de *méchant.* Il calcule ensuite la fonction potentielle associée à cette CSD et la compare à chacune des fonctions potentielles associées aux zones de sens de l'espace sémantique. Les calculs ont été menés pour les 40 noms les plus fréquemment utilisés comme recteur de *méchant* dans le corpus Frantext Catégorisé. Le tableau 3 présente quelques résultats:

ANTE			**POST**	
Sens intensifs Sens comportementaux	76% 24%	cheval	Sens intensifs	69%
Sens comportementaux	100%	regard	Sens comportementaux	100%
Sens intensifs	93%	homme	Sens intensifs	100%
Sens intensifs	99%	loup	Sens comportementaux Sens intensifs	53% 47%

Tableau 3 : Influence de la position de l'adjectif, quelques résultats sur l'adjectif *méchant*

Les résultats que nous obtenons montrent que le système est capable de repérer des changements de sens entre antéposition et postposition. Ce sont les sens intensifs qui obtiennent les scores les plus élevés dans quasiment tous les cas en antéposition. Le système est donc capable de rendre compte du fait qu'on trouve en antéposition les valeurs de sens qui ont la plus grande extension. Ici c'est clairement la valeur intensive, très générale, qui a la plus grande extension. Elle peut s'appliquer à n'importe quoi, alors que les deux autres valeurs ne s'appliquent qu'à des noms animés ou considérés comme tels. Le système repère en outre que, si en antéposition les sens intensifs sont omniprésents, il y a des noms pour lesquels ils ne s'imposent pas forcément. *Cheval, couleur, eau, espèce, farce, matin, mot, nature, parole, taureau, terre* donnent à *méchant* en antéposition tantôt une valeur générale, tantôt une valeur comportementale. Le contexte permet souvent de trancher entre les deux valeurs, mais ce n'est pas toujours très clair. On a ainsi une ambiguïté dans *un méchant cheval*, qui peut désigner selon les cas un cheval maigre, faible ou un cheval agressif. Nous avons pu vérifier que la fonction potentielle associée à la classe [ANTE] de *cheval* en tant que recteur de *méchant* présente deux sommets l'un couvrant la zone intensive et l'autre la zone comportementale. Cette ambiguïté disparaît en postposition, et le système en rend compte. Ainsi c*heval*, *couleur, dent, eau, espèce, farce, maison, mot, nature, parole, part, société, taureau, terre* acceptent aussi bien une valeur intensive que comportementale en antéposition, mais ne sélectionnent que la valeur comportementale ou psychologique en postposition. Pour les noms *bête, bois, bruit, chemin, chose, corps, rire et voix* le changement de sens est encore plus radical, puisqu'en antéposition *méchant* est exclusivement intensif, alors qu'en postposition il devient comportemental ou psychologique. Enfin, le système est aussi capable de repérer les noms pour lesquels on ne repère pas de changement de sens lors du passage de l'antéposition à la postposition. C'est le cas ici de *coup, part, regard* et *vérité*. Les erreurs rencontrées sur des noms comme *homme* ou *enfant* (calcul d'une valeur intensive en postposition) montrent ici encore les limites de la synonymie comme description du sens. Le calcul d'une valeur générale en postposition repose sur les hautes fréquences de cooccurrences de ces noms avec les adjectifs *maigre, faible, pauvre* et *petit.* Or la synonymie entre *méchant* et ces adjectifs n'est plus valable dans le contexte des noms considérés ici, en présence desquels *méchant* se colore plutôt d'une valeur comportementale ou psychologique. Le calcul d'un sens intensif en postposition pour *cheval* montre par ailleurs que ces relations de synonymie, en plus d'être partielles, ne sont valables qu'en antéposition. Là encore, une perspective de résolution du problème repose sur notre méthode d'exploration du graphe adjectival global, qui devrait permettre de repérer automatiquement les sens intensifs d'un adjectif, ceux qui ne sont valables qu'en antéposition.

8 Conclusion

Les analyses détaillées que nous avons menées montrent que les outils informatiques développés sont très prometteurs. Les différentes étapes dans la réalisation, et l'analyse des résultats obtenus à chaque pas, ont mis au jour différents problèmes, que nous devrons résoudre pour aller plus avant. L'utilisation des cliques du graphe de synonymie s'est avérée fort judicieuse, tant pour la construction des espaces sémantiques que pour le calcul du sens proprement dit. Le recouvrement de l'espace sémantique par les cliques contrebalance le fait que la relation de synonymie est une relation partielle, non transitive et peut dépendre de la position de l'adjectif, ce qui cause cependant encore quelques problèmes non résolus. Les classes distributionnelles ont montré leur efficacité pour la prise en compte de l'influence du nom recteur, et de sa position relativement à l'adjectif. Elles constituent une première étape vers une prise en charge des différences entre polysémie nominale et polysémie verbo-adjectivale. Jacquet (2006) dans son travail sur la polysémie verbale, arrive en effet à une conclusion similaire: « On pourrait envisager d'utiliser les CSD pour la désambiguïsation des noms en tant que tels. Cela reviendrait à dire que *bureau* dans l'énoncé *travailler sur le bureau* prend le sens de la classe (*bureau, table, chaise*) que l'on pourrait nommer '*meuble*'. Alors que dans *entrer dans le bureau*, *bureau* prend le sens de la classe (*bureau, cuisine, salon*) que l'on pourrait nommer '*pièce'*. » Le travail décrit ici ne constitue cependant que la première étape dans le processus de validation du modèle utilisé. Nous devons maintenant mener des évaluations massives sur un échantillon beaucoup plus large d'adjectifs ambigus. Il faudra, dans ce cadre, réfléchir au mode de représentation du sens utilisé. On peut chercher à trouver un ensemble de synonymes adéquats pour un adjectif dans un syntagme donné, ou encore réfléchir à une méthode automatique de partitionnement de l'espace sémantique par des outils géométriques (repérer les zones denses en cliques dans l'espace sémantique). Il faudra aussi tenir compte des erreurs inhérentes à l'analyse syntaxique automatique, par exemple dans le répérage du nom recteur de l'adjectif. Ce travail laisse cependant entrevoir tout l'intérêt de l'utilisation des mathématiques du continu en traitement automatique des langues.

Références

BOUILLON P.,VIEGAS E (1999). The description of adjectives for natural languge processing : theoretical and applied perspectives, in *Atelier thématique sur la description des adjectifs pour les traitements informatiques,* Institut d'études scientifiques de Cargëse.

FRANÇOIS J., VICTORRI B. et MANGUIN J.-L. (2002). Polysémie adjectivale et synonymie : l'éventail des sens de curieux, in *La polysémie,* Soutet, Olivier (Eds). Paris : PUS.

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

GOES J. (1999). *L'adjectif. Entre nom et verbe*. Paris/Bruxelles : Duculot.

JACQUET G. (2006), *Polysémie verbale et calcul du sens*; thèse de doctorat de l'EHESS.

JACQUET G., VENANT F. (2005). Construction automatique de classes de sélection distributionnelles, in *Actes du colloque TALN'05*, Dourdan.

PLOUX S., VICTORRI B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Traitement automatique des langues* 39 (1).

RASKIN V., NIRENBURG S. (1996.) Adjectival modification in text meaning representation. Proceedings of *COLING '96.* Copenhagen

VENANT F. (2006), *Représentation et calcul dynamique du sens : exploration du lexique adjectival du français*, thèse de doctorat de l'EHESS.