Polysémie et calcul du sens.

Fabienne Venant

Polysémie et calcul du sens.

Fabienne Venant
LaTTICe- ENS, 1 rue Maurice Arnoux-92120 Montrouge-France
fabienne.venant@ens.fr

Abstract

Polysemy is a pervasive phenomenon in natural languages, but it remains a vexing issue for natural language computing. In order to deal with this problem, Victorri and Fuchs (1996) proposed the model of *dynamical construction of meaning*. Each polysemic unit is associated to a semantic space. The meaning of the unit in a given sentence corresponds to a more or less restricted area of the semantic space, resulting from a dynamical interaction with all other units of the sentence. Ploux and Victorri (1998) designed a software, called Visusyn, allowing an automatic construction of the semantic space of a polysemic unity. The algorithm is based on the analysis of a large graph of synonyms (www.unicaen.crisco.fr). Visusyn was extended to take into account the data from one corpus (base Frantext categorisée). The performance of Visusyn was compared to that of french speakers in a disambiguation task. The results of this experiment strongly incited us to start a theoretical mathematic and informatic study of the graph of synonyms. This study will be crucial for a better understanding of french lexical organisation.

Résumé

La polysémie est un phénomène omniprésent dans le langage, mais il reste problématique dans le cadre du traitement automatique des langues. Nous nous proposons d'aborder ce problème à l'aide du modèle de *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 (1998) ont développé Visusyn, un logiciel permettant de construire automatiquement l'espace sémantique d'une unité polysémique. L'algorithme repose sur l'analyse d'un grand graphe de synonymie (www.unicaen.crisco.fr). Nous avons étendu Visusyn afin qu'il prenne en compte les données issues d'un corpus (Frantext catégorisée). Nous avons comparé les performances de Visusyn avec celles de locuteurs du français dans une tâche de désambiguïsation. Les résultats de cette expérience nous ont encouragé à entreprendre une étude théorique mathématique et informatique du graphe de synonymie. Celle-ci devrait s'avérer cruciale pour une meilleure compréhension de l'organisation du lexique français.

Mots clés : polysémie, calcul du sens, espace sémantique, désambiguïsation, grands graphes, small world.

1. La polysémie en linguistique computationnelle.

La plupart des unités lexicales que nous utilisons ont plusieurs sens. Loin de nous gêner pour communiquer, ce phénomène, appelé polysémie, est au contraire source de richesse et de souplesse dans les langues. Nous sommes habitués à manier les indices contextuels et nous comprenons instantanément le sens de n'importe quel mot polysémique dans n'importe quel énoncé. Pourtant dès que l'on veut automatiser une telle performance, la polysémie devient un véritable problème et elle donne bien du souci aux chercheurs en traitement automatique du langage. La prise en compte de la polysémie en TAL se traduit par la question suivante : « comment associer automatiquement un sens à un mot dans un énoncé donné ? ». La tâche s'effectue donc en deux étapes: d'abord déterminer tous les sens possibles pour chaque mot susceptible d'être désambiguïsé et ensuite déterminer quel sens est le bon en contexte. Les ordinateurs vont utiliser les mêmes indices que nous, à savoir le contexte. Ce qui leur manque c'est toute notre connaissance du lexique et de son organisation. C'est là tout l'enjeu des tâches de désambiguïsation. Il y a deux pistes suivies majoritairement: l'une consiste à utiliser les distinctions de sens présentes dans les dictionnaires, l'autre à utiliser des méthodes statistiques pour repérer des patterns de cooccurrences des mots en contexte. En 1998, *Computational Linguistics* a spécialement consacré un article à la question de la désambiguïsation sémantique (Ide et Véronis, 1998).

Dans les années 80 les ressources lexicales à grande échelle (dictionnaires électroniques, glossaires, thésaurus, ontologies…) se sont développées et beaucoup de travaux ont utilisé les divisions de sens fournies par ces outils. L'idée est que le sens le plus probable pour une occurrence d'un mot donné est celui qui va maximiser une certaine relation d'affinité avec le contexte de cette occurrence. Lesk (1986) a créé une méthode permettant de relier des définitions si elles ont des mots en commun. La désambiguïsation d'un mot en contexte se fait en choisissant pour lui et les mots qui l'entourent les définitions qui se recoupent le plus. Cette méthode est très sensible à la présence ou non d'un mot dans une définition et pose problème en cas de définitions lapidaires. D'autres études ont donc cherché à l'améliorer en utilisant d'autres indices. Walker (1987) pense à utiliser des codes de sujet (ensemble de primitives servant à classifier les sens d'un mot par sujet) . Guthrie et al. (1991) complètent la méthode de Lesk en imposant une correspondance entre codes de sujets dans un processus itératif. Ils calculent des "voisinages" de mots polysémiques en cherchant des cooccurrences dans les définitions de tous les mots qui partagent un même code de sujet. Wilks et al. (1990) utilisent les fréquences de cooccurrences pour les mots dans les textes des définitions et en déduisent des degrés d'affinités entre mots. Ils les utilisent ensuite dans une méthode vectorielle qui relie chaque mot et son contexte. Veronis et Ide (1990) ont prolongé la méthode de Lesk en créant un réseau de neurones à partir des définitions du Collins English Dictionary: chaque mot est relié à ses sens, eux-mêmes reliés aux mots de leurs définitions, eux même reliés à leur sens. Wilks et al. (1993) ont beaucoup réfléchi à la façon d'utiliser de façon optimale les ressources électroniques pour identifier les sens des mots polysémiques.

Plus récemment on a vu se développer des méthodes de désambiguïsation sémantique sur corpus. Il s'agit d'analyser les mots qui cooccurrent avec les mots polysémiques sur des corpus à grande échelle. Ces systèmes s'entraînent à modéliser le sens de chaque mot, en fonction de leur contexte, à partir de corpus d'exemples sémantiquement étiquetés (de 50 à 100 phrases). Ils choisissent ensuite le sens le plus adéquat pour une nouvelle occurrence d'un mot dans le texte à traiter. L'adéquation d'un sens est calculée à partir d'une mesure de similarité entre les caractéristiques des sens modélisés et celles du contexte de l'occurrence considérée. Les travaux de ce genre sont très nombreux. Kilgariff (1997) a organisé une conférence, Senseval, dans le but de faire le point et de comparer les différents travaux de désambiguïsation à partir

de corpus. Il s'agissait de désambiguïser un ensemble de 35 mots polysémiques (noms, verbes, adjectifs) préalablement choisis. Pour chacun d'entre eux, un corpus d'occurrences était fourni. Les systèmes devaient déterminer le (ou les) sens (éventuellement accompagnés d'une probabilité) de chaque occurrence. Les sens étaient ceux de la base de données Hector. Les résultats fournis étaient ensuite comparés à ceux donnés par des juges humains sur le même corpus. Les meilleures performances ont été de 80% de réussite pour les noms, 70% pour les verbes et 75% pour les adjectifs. Les systèmes utilisant un corpus d'apprentissage ont globalement mieux réussi que les autres. (voir Kilgarrif et Rozenzweig, 2000, pour plus de détails).

Une des difficultés de la tâche vient de l'inventaire des sens lui-même. La plupart des travaux réalisés reposent sur des dictionnaires traditionnels, ou des ressources électroniques comme Wordnet, qui ne diffèrent pas énormément en termes de division de sens. Le problème est que les dictionnaires ont été réalisés pour un usage humain et non pas automatique. Ils manquent donc d'informations pragmatiques utiles à la désambiguïsation. D'autre part, l'inconsistance des dictionnaires est bien connue des lexicographes (Kilgariff , 1994). Véronis (2001) pense qu'on ne pourra pas progresser en désambiguïsation sémantique tant que les dictionnaires n'incluront pas dans leurs définitions des critères distributionnels ou des indices de surface (syntaxes, collocations ,…). C'est pourquoi au sein de son équipe, Reymond travaille à la réalisation d'un dictionnaire "distributionnel" spécialement adapté au problème de la désambiguïsation par des machines (Reymond, 2001). Il s'agit d'organiser les mots en lexies possédant des propriétés distributionnelles cohérentes. Audibert travaille, à partir de ce dictionnaire, à étudier les différents critères de désambiguïsation (cooccurrence, n-grammes, information sur le domaine, synonymes des mots en cooccurrence….) (Audibert, 2002-2003).

Une autre voie de recherche suivie par Véronis (2003), toujours dans l'idée de pallier aux insuffisances des dictionnaires classiques en matière de discrimination des sens, est l'utilisation d'un graphe de cooccurrence. Il s'agit de déterminer automatiquement les différents usages d'un mot dans une base textuelle (en l'occurrence google). L'algorithme est basé sur la recherche des zones de forte densité du graphe de cooccurrences et permet contrairement aux méthodes classiques d'analyse textuelle (comme les vecteurs de mots), d'isoler des usages très peu fréquents. Jean Véronis met ici en application le conseil de Wittgenstein: "Don't look for the meaning, but for the use".

Une autre solution est de travailler sur des relations paradigmatiques entre mots (synonymie, antonymie,…). Comme le remarquent Edmond et Hirst (2002) un mot peut exprimer une myriade d'implications, de connotations en plus de son sens dans les dictionnaires. Un mot a des synonymes (il s'agit ici de relation de synonymie partielle) qui diffèrent de lui dans ces nuances de sens. Ils cherchent à développer un modèle computationnel de la connaissance lexicale qui rende compte adéquatement de la "presque synonymie" et qui dans une tâche de traduction automatique puisse choisir le bon mot, celui qui va rendre compte de la nuance de sens exacte, dans un contexte donné. L'idée étant de pouvoir rendre compte des sens indirects, flous ou dépendant du contexte ignorés des systèmes actuels.

A la lumière de tous ces travaux, on peut s'étonner avec Véronis (2001) du fait que la pertinence cognitive ne soit jamais recherchée. Une expérience qu'il a menée montre que les humains eux-mêmes ont de piètres performances quand il s'agit d'associer un sens d'un dictionnaire à une occurrence d'un mot dans un énoncé. Mis à part Edmond et Hirst, on s'interroge peu sur le fait qu'une occurrence d'un mot puisse jouer sur plusieurs sens possibles sans qu'on puisse trancher entre les deux. Ce phénomène qu'on appelle indétermination est pourtant au cœur même de l'expressivité d'une langue. Enfin aucune des méthodes actuelles ne s'interroge réellement sur l'organisation du lexique. Même les méthodes basées sur des

calculs de similarités ne cherchent pas à représenter les distances sémantiques entre sens et ne parviennent pas à organiser correctement les sens obtenus.

Le modèle que nous utilisons s'appuie sur un espace sémantique où sont organisés les différents sens d'un mot. La polysémie est le mécanisme central dont nous cherchons à rendre compte. Le calcul du sens d'un énoncé est un processus dynamique au cours duquel les sens des différents mots s'influencent mutuellement et qui aboutit simultanément à la détermination du sens de chacun des mots et à un sens global pour la phrase. C'est le principe de la compositionnalité gestaltiste défini par Victorri et Fuchs (1996). Ce modèle a été implémenté en utilisant deux ressources lexicales : un dictionnaire électronique de synonymes et un grand corpus.

2. Calcul dynamique du sens

Le principe de compositionnalité gestaltiste implique une modélisation dans le cadre des systèmes dynamiques. Cela permet d'éviter le cercle vicieux du au fait que la plupart des unités sont polysémiques, et que pour calculer le sens de chacune d'elles on a besoin de connaître les sens des autres, et réciproquement.

La donnée d'une dynamique sur un espace revient à spécifier les contraintes qui s'exercent en chaque point de cet espace et permet d'obtenir les points de stabilisation qui correspondent aux solutions du problème. Il s'agit donc ici d'associer à chaque unité linguistique un espace, appelé espace sémantique, muni d'une structure mathématique précise où le sens de l'unité dans chacun de ses emplois est représenté par une région de l'espace.

Les unités cotextuelles définissent une « fonction potentielle » sur l'espace sémantique. Les valeurs du potentiel inférieures à un certain seuil déterminent une région de l'espace sémantique qui représente le sens de l'unité dans l'énoncé considéré

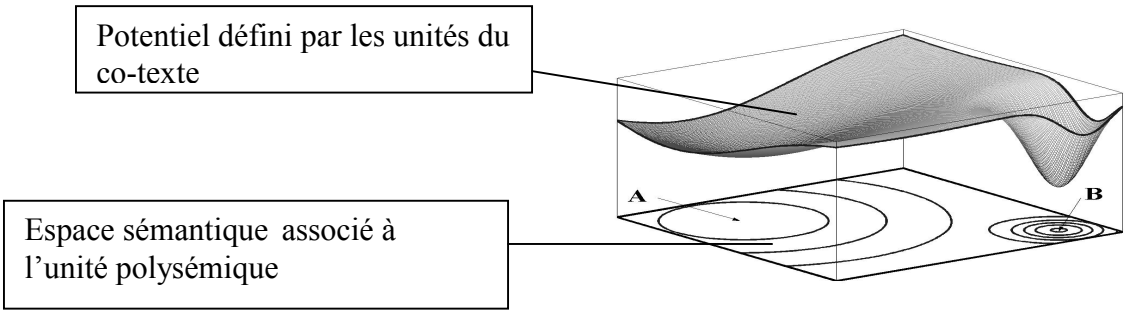


Figure1 : représentation d'une fonction potentielle sur un espace sémantique bidimensionnel

3.Visusyn

Pour construire et représenter de façon automatique l'espace sémantique associé à un mot, nous utilisons le logiciel VISUSYN développé par Ploux et Victorri (1998). Pour déterminer automatiquement les paramètres de l'espace sémantique, Visusyn utilise un algorithme basé sur l'analyse d'un graphe de synonymie. Ce graphe nous est fourni par le Dictionnaire Electronique des Synonymes (D.E.S.) du laboratoire CRISCO (www.crisco.unicaen.fr)

Prenons un exemple : on s'intéresse ici à l'adjectif *sec*. Le D. E. S. fournit la liste des synonymes de *sec* (au nombre de 63) . Il en construit le graphe : les sommets sont *sec* et ses synonymes, deux unités sont en relation si elles sont synonymes. On trouvera en Figure 2 un extrait du graphe de synonymie associé à *sec*.

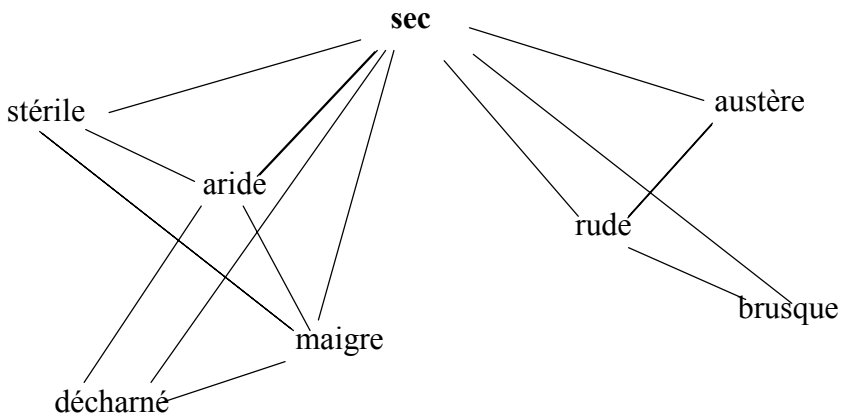


Figure 2 : un extrait du graphe de synonymie de *sec*.

L'idée sous jacente à la construction de l'espace sémantique est semblable à celle développée par Edmond et Hirst (2002). Un simple synonyme n'est généralement pas suffisant pour définir un sens précis d'une unité. On voit ici que *aride* est à la fois synonyme de *décharné* et de *stérile*, ce qui correspond à deux sens distincts de *sec* ; l'un lié à la maigreur l'autre au caractère improductif. Or les points de notre espace sémantique doivent correspondre à des sens précis de l'unité. C'est pourquoi nous avons recours à la notion de clique. Une clique est un sous-graphe complet maximal, c'est à dire un ensemble de sommets, le plus grand possible, reliés deux à deux. La portion de graphe que nous avons représentée ici contient 4 cliques : *Aride-décharné-maigre-sec; Aride-maigre-stérile-sec*; *Austère-rude-sec* et *Brusque-rude-sec*. Chaque clique correspond à une nuance possible de sens pour *sec*.

Le D. E. S. fournit la liste des cliques du graphe correspondant à une unité donnée et ce sont ces cliques que nous allons représenter. Notre espace sémantique est une projection en deux dimensions du nuage formé par les cliques dans l'espace multidimensionnel engendré par les synonymes de l'unité lexicale considérée. Il est muni de la métrique du χ^2. C'est elle qui s'est en effet avérée la plus efficace pour obtenir une représentation respectant la notion intuitive de proximité entre sens. On trouvera une représentation de l'espace sémantique de *sec* en figure 3.

4. Désambiguïsation d'un adjectif

Outre la construction de l'espace sémantique associé à une unité, Visusyn permet d'obtenir la visualisation d'une zone associée à chaque synonyme de l'unité étudiée. Nous avons étendu ce logiciel afin qu'ils puissent prendre en compte des données issues d'un corpus pour visualiser la zone dans laquelle un nom contraint un adjectif à prendre son sens. Nous illustrons ici ces deux propriétés dans le cas de l'adjectif *sec* et nous montrons ensuite comment nous les avons utilisées dans une expérience de désambiguïsation de l'adjectif *sec* en rection nominale.

4.1 Sémantique de l'adjectif sec.

Sec est un adjectif très polysémique mais dont on peut regrouper les sens en six acceptions principales :

(1) qui manque d'eau : *du sable sec*

(2) maigre, décharné : *un homme grand et sec*

(3) stérile, improductif : *rester sec aux questions du professeur*

(4) qui manque de sensibilité, qui ne se laisse pas attendrir, égoïste : *un cœur sec*

(5) bref, abrupt, qui manque de douceur *: un coup sec*

(6) seul : *un atout sec.*

Bien que ces sens soient très différents, ils peuvent être reliés les uns aux autres par une "ressemblance de famille" à la Wittgenstein. Les sens (1), (2) et (3) se rejoignent lorsque *sec* qualifie de la végétation. De même les sens (3) et (4) sont liés : une personne sèche au sens d'égoïste est quelqu'un de stérile en termes d'empathie et de don de soi. On sent aussi une relation entre le sens (5), qui s'applique à des événements, et le sens (4) qui caractérise un comportement mal dégrossi.

Ce sont toutes ces proximités de sens dont notre représentation doit rendre compte et on peut voir sur la figure 3 que le résultat est plutôt satisfaisant.

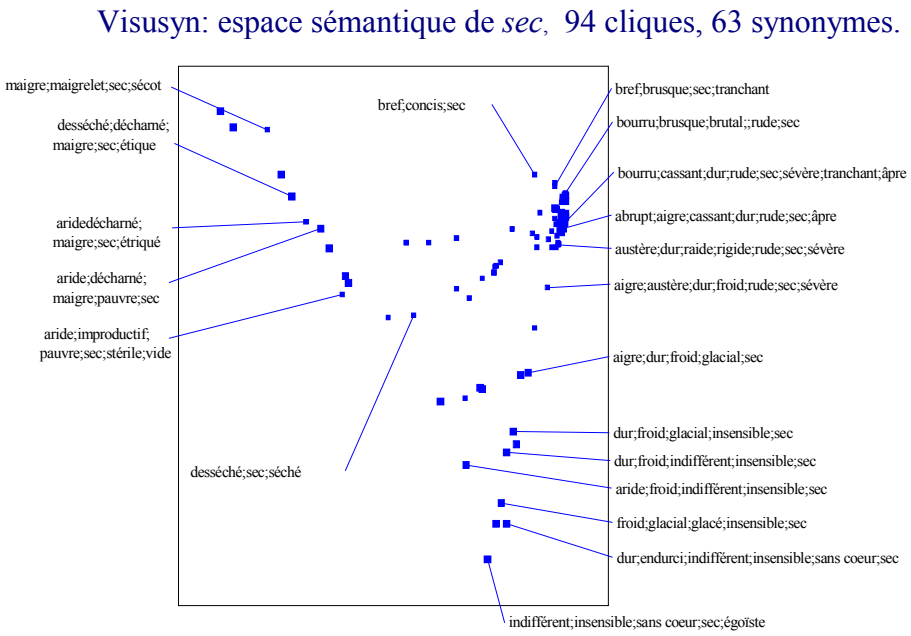


Figure 3 : Espace sémantique associé à l'adjectif *sec*

4.2 Zone de pertinence associée à un synonyme.

A chaque synonyme du mot vedette, on associe une fonction dont les bassins représentent de façon plus précise la zone de sens occupée par ce synonyme. Cette fonction permet de visualiser la région de l'espace sémantique dans laquelle la relation de synonymie entre le mot vedette et le synonyme considéré est pertinente. A titre d'exemple, on trouvera en figure 4 et 5 les zones de pertinence de *brusque* et *aride*

Figure 4.: Zone de pertinence *de brusque*

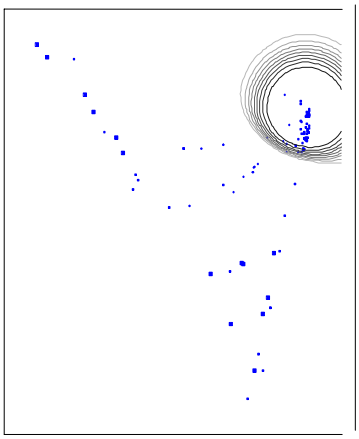
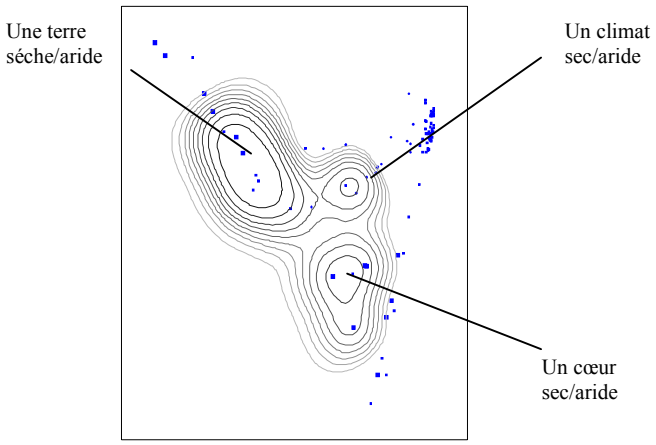


Figure 5: zone de pertinence de *aride.*



4.3 Potentiel désambiguïsateur du nom régissant.

Ce qui permet de désambiguïser un adjectif c'est essentiellement la présence du nom régissant. Notre modèle permet de rendre compte de ce phénomène. Dans un premier temps, on relève tous les noms en cooccurrence avec l'adjectif considéré (le corpus utilisé est la base Frantext catégorisée) et on calcule le nombre de cooccurrences de chacun de ces noms avec chacun des synonymes de l'adjectif vedette. On obtient par exemple pour *sec* une liste de 126 unités (*air*, *arbre*, *bois*, *boue*, *bras*, *chambre*, *chemin*, *cheveu*, *chose*, *claquement*, *cœur*, *coin*, *corps*, *cou*, *coup*, *doigt*, *éclat*, *esprit*, *façon*, *femme*, *feuille*, *figure*, *fleur*, *foin*, *fromage*, *geste*, *gorge*, *herbe*, *homme*, *jambe*, *jardin*, *lèvre*,...). A partir de ces données, Visusyn calcule le degré d'affinité d'un nom avec une clique : Plus le nombre de cooccurrences du nom avec les éléments d'une clique est grand, plus son degré d'affinité avec cette clique est élevé. Voici par exemple le nombre de cooccurrences de *coup* avec les adjectifs suivants :

bref : 67, brusque : 48, tranchant : 0, sec : 173, *maigre : 0, maigrelet* : *0*, *sécot : 0.*

Le nom coup est très compatible avec l'adjectif *bref*, un peu moins avec *brusque* et pas du tout avec *tranchant, maigre, maigrelet* et *sécot*. Son degré d'affinité sera donc assez élevé avec la clique *sec; bref; brusque; tranchant* et assez faible avec *sec ; maigre; maigrelet; sécot.*

Le calcul exact prend en compte les fréquences relatives des adjectifs dans le corpus et ne sera pas détaillé ici. (voir Ploux et Victorri, 1998). Visusyn utilise ensuite ce degré d'affinité pour construire une fonction potentielle associée à chaque nom. Cette fonction permet de visualiser la zone de sens pertinente dans le contexte du nom considéré. On peut voir en figure 6 que les résultats obtenus pour des mots comme *fleur* et *ton* sont très bons. La fonction associée à fleur possède un unique bassin très étroit. *Sec* a dans ce contexte un sens très précis qui est celui de manque d'eau. Pour *ton*, on obtient un bassin très large qui couvre la presque totalité de la partie droite de l'espace sémantique. On a dans le cas de *ton sec* une indétermination. Un *ton sec* l'est aussi bien d'un point de vue psychologique (en bas à droite de l'espace) que physique (en haut à droite). Il existe cependant des noms comme *lit* et *visage* pour lesquels les potentiels désambiguïsateurs ne correspondent pas aux contraintes réelles qu'ils exercent sur la sémantique de *sec.* C'est pourquoi il nous a paru nécessaire d'étudier plus en détails les résultats obtenus afin de dégager les forces et les faiblesses de la méthode.

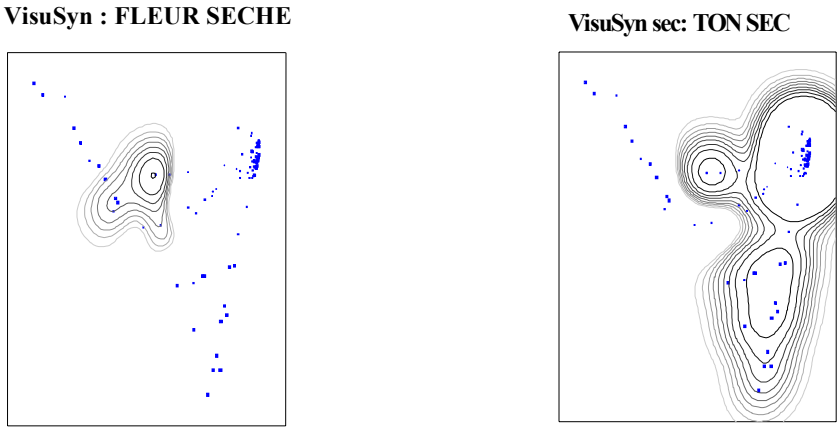


Figure 6: Potentiels désambiguisateurs des noms *fleur* et *ton*

4.4 Evaluation

Nous avons voulu évaluer la pertinence de nos calculs en les comparant à ceux obtenus par des locuteurs du français. Nous avons donc conçu une tâche de désambiguïsation réalisable à la fois par Visusyn et par les sujets. Il s'agit de sélectionner parmi les 5 synonymes proposés, *brusque, décharné, desséché, stérile, glacial*, celui ou ceux qui décrit le mieux le sens de *sec* en présence d'un nom donné.

Pour ce faire Visusyn compare les fonctions potentielles du nom et de l'adjectif et attribue une note entre 1 et 5 à l'adjectif. Plus les zones de sens de l'adjectif et du nom sont proches, plus la note est proche de 1. On trouvera en page 9 le tableau des résultats. Avec un taux de réussite globale de 79%, et alors que nous n'utilisons pas de corpus d'entraînement, nous sommes au dessus du taux de réussite obtenu pour les adjectifs dans Senseval. Cependant la comparaison reste difficile à faire puisque nous ne travaillons pas sur la même langue, ni sur le même corpus. Enfin nous n'utilisons pas les sens définis dans un dictionnaire si bien que nous ne pouvons pas nous comparer non plus aux résultats obtenus lors de l'évaluation Romanseval réalisée sur le français à partir des sens définis dans le petit Larousse 95, et dont les résultats ont été extrêmement modestes (Segond, 2000).

Pour presque 63% des cas considérés, le taux de réussite est supérieur à 70%, ce qui veut dire que pour ces noms le calcul automatique du sens aboutit à la sélection d'un sens de *sec* valable dans le contexte du nom considéré.

Ainsi nous pouvons calculer correctement le sens de *mouvement sec* (***brusque***), *coup sec* (***brusque***), *main sèche* (***desséchée***), *corps sec* (***décharné***), *terre sèche* (***desséchée*** / ***stérile***), *ton sec* (***brusque***), *éclair sec* (***brusque***), *torrent sec* (***desséché***), *cou sec* (***décharné*** / ***desséché***), *manières sèches* (***brusques***), *sol sec* (***desséché***), *arbre sec* (***décharné / desséché***).

Dans la majorités des cas, nous pouvons déterminer s'il s'agit d'un synonyme parfait ou approximatif (attribution correcte de la note un ou deux) : *mouvement sec* et *mouvement brusque, coup sec* et *coup brusque, corps sec* et *corps décharné, terre sèche* et *terre desséchée, éclair sec* et *éclair brusque, cou sec* et *cou décharné* sont parfaitement synonymes. En revanche pour *main sèche* et *main desséchée*, *terre sèche* et *terre stérile, cou sec* et *cou desséché*, le remplacement de *sec* par son synonyme provoquera un léger changement de sens. Tout cela notre calcul le prédit parfaitement.

Pour les deux noms *coup* et *éclair* la réussite est même parfaite (100%), c'est à dire que non seulement le calcul sélectionne le bon sens de *sec* en présence du nom, mais avec la même précision que les sujets (note 1), et de plus il rejette aussi nettement les sens qui ne conviennent pas.

Pour les noms *terre, arbre et cou* on a obtenu les deux sens possibles sélectionnés par les sujets (*terre desséchée / stérile*, *arbre ou cou décharné / desséché*). Cependant certains cas d'indétermination nous échappent (*ton brusque / glacial, manière brusques / glaciales,*).

L'analyse détaillée des erreurs nous a permis de dégager plusieurs voies de travail en vue d'améliorer notre système.

nom	adjectif	note Visusyn	note sujets	réussite(%)		nom	adjectif	note visusyn	note sujets	réussite(%)
mouvement	brusque	1	1	100		**torrent**	brusque	4	4	100
	décharné	4	4	100			décharné	4	4	100
	desséché	4	4	100			desséché	2	1	80
	stérile	4	4	100			stérile	4	4	100
	glacial	4	3	80			glacial	4	4	100
			moyenne	**96**					**moyenne**	**96**
vent	brusque	1	4	0		**cou**	brusque	4	4	100
	décharné	4	4	100			décharné	1	1	100
	desséché	4	2	0			desséché	2	2	100
	stérile	4	3	80			stérile	3	4	80
	glacial	4	4	100			glacial	4	4	100
			moyenne	**56**					**moyenne**	**96**
coup	brusque	1	1	100		**manière**	brusque	1	2	80
	décharné	4	4	100			décharné	4	4	100
	desséché	4	4	100			desséché	4	4	100
	stérile	4	4	100			stérile	4	4	100
	glacial	4	4	100			glacial	4	2	0
			moyenne	**100**					**moyenne**	**76**
main	brusque	3	4	80		**sol**	brusque	4	4	100
	décharné	2	3	20			décharné	3	4	80
	desséché	2	2	100			desséché	2	1	80
	stérile	3	4	80			stérile	3	2	20
	glacial	3	4	80			glacial	3	4	80
			moyenne	**72**					**moyenne**	**72**
visage	brusque	3	4	80		**arbre**	brusque	4	4	100
	décharné	3	2	20			décharné	1	2	80
	desséché	3	2	20			desséché	2	1	80
	stérile	4	4	100			stérile	4	4	100
	glacial	4	3	80			glacial	4	4	100
			moyenne	**60**					**moyenne**	**92**
corps	brusque	3	4	80		**souffle**	brusque	1	3	0
	décharné	1	1	100			décharné	4	4	100
	desséché	2	1	80			desséché	4	1	0
	stérile	3	4	80			stérile	4	4	100
	glacial	3	4	80			glacial	3	3	100
			moyenne	**84**					**moyenne**	**60**
terre	brusque	4	4	100		**boue**	brusque	4	4	100
	décharné	3	3	100			décharné	4	4	100
	desséché	1	1	100			desséché	4	1	0
	stérile	2	2	100			stérile	4	4	100
	glacial	3	4	80			glacial	1	4	0
			moyenne	**96**					**moyenne**	**60**
éclat	brusque	1	4	0		**lit**	brusque	3	4	80
	décharné	4	4	100			décharné	4	4	100
	desséché	4	4	100			desséché	4	1	0
	stérile	4	4	100			stérile	4	4	100
	glacial	4	4	100			glacial	3	4	80
			moyenne	**80**					**moyenne**	**72**
ton	brusque	1	2	80		**fleur**	brusque	4	4	100
	décharné	4	4	100			décharné	4	3	80
	desséché	4	4	100			desséché	3	1	0
	stérile	4	4	100			stérile	2	4	0
	glacial	4	2	0			glacial	4	4	100
			moyenne	**76**					**moyenne**	**56**
						RECAPITULATIF				
éclair	brusque	1	1	100		**taux moyen de**	arbre 92%	éclat 80%	sol 72%	visage 60%
	décharné	4	4	100		**réussite: 79%**	boue 60%	fleur 56%	souffle60%	
	desséché	4	4	100			corps 84%	lit 72%	terre 96%	
	stérile	4	4	100			cou 96%	main 72%	ton 76%	
	glacial	4	4	100			coup 100%	manière76%	torrent96%	
			moyenne	**100**			éclair 100%	mvt 96%	vent 56%	

Nous travaillons actuellement à agrandir le corpus étudié. Le fait qu'on ait eu accès à un nombre limité de textes, en majorité très littéraires, explique que certains sens d'un usage plus quotidien échappent à notre calcul. Parallèlement à l'élargissement du corpus, il nous faut perfectionner notre mode de représentation et de calcul. Nous devons donner plus de poids aux cliques centrales, qui, bien que possédant peu de synonymes, représentent des sens très importants du mot vedette et qui doivent donc prendre plus de poids dans les calculs. Il faudra étudier en détails quelle est la façon la plus efficace de rééquilibrer notre représentation. Un autre biais dans le calcul est du au fait que la présence de cliques ayant un fort degré d'affinité avec un nom donné peut être contrebalancée par le voisinage de cliques ayant un degré d'affinité très faible. Une piste possible pour la résolution de ce problème est d'augmenter le nombre de dimensions de l'espace de représentation. Il semblerait intéressant aussi d'affiner notre construction de l'espace sémantique en utilisant d'autres relations paradigmatiques comme l'antonymie, voire en croisant divers indices de proximité entre synonymes comme l'ont fait Inkpen et Hirst (2003).

Enfin, le problème le plus important que nous ayons rencontré vient de ce que deux synonymes de *sec* peuvent se trouver en cooccurrence avec un même nom sans pour autant être dans ce cas synonymes entre eux. Visusyn considère par exemple que *boue sèche* et *boue glaciale* sont synonymes, du fait que *boue* est à la fois très compatible avec *glacial* et avec *sec*. Notons que l'on rencontre le même phénomène avec *temps* : *temps sec* n'est pas synonymes de *temps glacial*. Le problème se pose ici d'abord parce que *sec* peut prendre des sens dépendant de domaines différents: les uns sont des sens physiques (*un arbre sec, du sable sec*), les autres psychologiques (*un cœur sec*), ensuite parce que parmi ses synonymes, certains comme *froid*, *glacé* et *glacial* peuvent aussi déployer leur sens dans les deux domaines (*une eau froide / un abord froid, une boisson glacée / un accueil glacé*, *un vent glacial / un sourire glacial*), enfin parce que *sec*, *froid*, *glacial*, *glacé* ne sont synonymes que dans leurs sens psychologiques mais que certains noms, aussi utilisés avec *sec*, peuvent se trouver en cooccurrence avec eux dans un sens physique qui échappe à la synonymie de *sec*. Notons d'autre part que si ce phénomène a une incidence notable sur nos calculs, c'est aussi parce que ces adjectifs partagent de nombreuses cliques. En effet notre méthode de calcul est robuste et lorsqu'un tel phénomène ne concerne qu'un nombre limité de cliques, il n'interfère pas dans les calculs.

C'est ce genre de problèmes qui nous ont convaincus qu'une étude théorique mathématique et informatique plus approfondie de la structure du graphe de synonymie était nécessaire pour progresser. Pour prendre en compte des interactions aussi subtiles que celle de *sec*, *glacial* et *froid* il faudrait pouvoir obtenir une visualisation plus globale du graphe de synonymie. Notre méthode de représentation à partir des cliques, dans laquelle les unités lexicales occupent des régions plus ou moins grandes suivant leur polysémie est assez efficace mais essentiellement locale. On ne peut visualiser le graphe au voisinage d'un sommet. Nous cherchons actuellement, en collaboration avec Bruno Gaume (Irit Toulouse) à définir des méthodes permettant de visualiser des parties plus importantes du graphe. En utilisant des structures plus « lâches » que les cliques, les « gangs », on peut obtenir des cartes du graphe à différentes échelles, de la plus locale à la plus globale. Nous voudrions construire ainsi une sorte d'atlas sémantique du français.

5. Conclusions et perspectives.

Les résultats de l'expérience sont très encourageants. Nous avons un taux de réussite de 79%. Nous pouvons raisonnablement considérer que, même sans résoudre immédiatement tous les problèmes théoriques, il va pouvoir atteindre les 90%. Outre les voies de recherche dégagées dans cet article, nous comptons beaucoup sur le travail de Jacquet (2003) sur la

désambiguïsation des verbes et l'influence des constructions verbales dans la construction du sens. Nous pouvons donc espérer proposer rapidement une méthode automatique et complète de calcul du sens sur des grands corpus.
D'autre part le travail entrepris sur les grands graphes devrait nous mener plus loin que l'avancée de nos travaux en modélisation de la polysémie. Les graphes sont de plus en plus utilisés en sémantique, notamment dans les études portant sur la connaissance lexicale. Les sommets de ces graphes représentent les mots d'une langue. Il existe plusieurs types de réseaux selon la relation sémantique utilisée pour définir les arcs du graphe. Celle ci peut être de type syntagmatique ou de cooccurrence : on construit un arc entre deux mots si on les trouve au voisinage d'un mot cible (Veronis, 2003). Elle peut être de type paradigmatique comme c'est le cas dans notre graphe de synonymie. Il peut s'agir d'une relation plus générale de proximité sémantique prenant en compte à la fois l'axe paradigmatique et l'axe syntagmatique (Gaume et al., 2002). On peut enfin imaginer de relier des mots sur des critères distributionnels, suivant les contextes qu'ils partagent, comme le fait Bourigault (2002). Ces graphes, tout comme la plupart des grands graphes de terrain (graphes sémantiques, réseaux géographiques, électriques, Internet…) partagent une topologie bien particulière (peu denses, structure hiérarchique et connectivité locale très forte) et on les appelle graphes de type « small world ».Les outils que nous allons mettre en place pourront donc être utilisés directement dans d'autres domaines des sciences sociales.
L'idée qui sous tend nos travaux est que ces graphes recèlent dans leur structure une information très riche mais difficile à utiliser directement. Notre méthode de géométrisation devrait permettre une meilleure compréhension des phénomènes sous jacents. On peut ainsi espérer en savoir plus sur la structure sémantique du lexique d'une langue et mettre au point des méthodes de navigation dans le lexique.

Références

Audibert. L. (2003). Etude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences. *In Actes TALN 2003*, pages 35-44.

Audibert. L. (2002). Etude des critères de désambiguïsation sémantique automatique : présentation et premiers résultats sur les cooccurrences. *In Actes de RECITAL (TALN) 2002*, pages 415-424.

Bourigault D (2002), Upery : un outil d'analyse distributionnelle étendue pour la construction

D'ontologies à partir de corpus. *In Actes de la 9ème conférence annuelle sur le Traitement*

Automatique des Langues (TALN 2002).

Edmonds P. and Hirst G. (2002). Nearsynonymy and lexical choice. *Computational Linguistics*, 28(2):105–144.

Gaume B., Duvignau K., Gasquet O. and Gineste M-D. (2002). Forms of Meaning, Meanings of Forms. *Journal of Experiment and Theoretical Artificial Intelligence*, 14(1): 61-74.

Guthrie J.A., Guthrie L., Wilks Y. and Aidinejad H. (19991). Subject-dependent co-occurrence and word sense disambiguation . In Morristown NJ: ACL, *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 146-152.

Ide N. et Véronis J (1998). Introduction to the special issue on word sense disambiguation : the state of the art. *Computational linguistic*, 24.I: 1:40.

Jacquet G. (2003). Polysémie verbale et construction syntaxique : étude sur le verbe jouer. *In Actes TALN 2003*, pages 469-479.

Kilgariff A (1994). The myth of completeness and some problems with consistency (the role of frequency in deciding what goes in the dictionary. *In Proceedings of the 6th International Congress on Lexicography, EURALEX'94*., pages 101-106.

Kilgariff A (1997). Evaluating word sense disambiguation programs: progress report. Information Technology Research Institute. Brighton:
Kilgarriff A., Rosenzweig J. (2000), English SENSEVAL: Report and Results, Actes de *2nd International Conference on Language Resources and Evaluation*, pp.1239-1244.

Inkpen D. Z. and Hirst G., Automatic sense disambiguation of the near-synonyms in a dictionary entry. *In Proceedings, 4th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2003)*,

Lesk M. (1986) .Automated Sense Disambiguation: How to Tell Pine Cone from an Ice Cream Cone. In New York : Association for Computing Machinery, *Proceedings of the 19986 SIGDOC Conférence,,* pages 24-26.

Ploux S. and Victorri B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires informatisés des synonymes. *TAL*, 39(1):161–182.

Ravin Y. and Leacock C. *Polysemy: Theoretical and Computational Approaches*. New York: Oxford University **Press.**

Reymond D (2001). Dictionnaires distributionnels et étiquetage lexical de corpus. *In Actes de la Conférence Traitement Automatique des Langues (RECITAL'2001)*, pages 24-33.
Segond, F. (2000). Framework and results for French. *Computers and the Humanities, 34*(1/2), 49-60 [special issue on Senseval].

Schütze H. (1992). Dimensions of meaning. In IEEE Computer Society press ,editors, *Proceedings of Supercomputing'92*, pages 787-796.

Schütze H. (1993). Word space. In Hanson S., Cowan J. and Giles C. L., editors, *Advances in Neural Information Processing Systems*. Morgan Kauffmann.

Véronis J. (2001). Sense tagging: does it make sense?, *Corpus Linguistics'2001* .

Véronis J. (2003). Cartographie lexicale pour la recherche d'information. *In Actes TALN 2003*, pages 265-275.

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

Victorri B. (2002), Espaces sémantiques et représentation du sens,. *Textualités et nouvelles technologies*, éc/artS, 3.

Walker D.E. (1987). Knowledge resource tools for accessing large text files. In S. Nirenburg, (ed.), *Machine Translation*. Cambridge University Press.

Véronis J. and Ide N (1990). Word Sense Disambiguation with Very Large Neural Network Extracted from Machine Readable Dictionaries. *13th International Conference on Computational Linguistics*, *COLING'90*, vol. 2, 389-394.

Wilks Y. A. and Fass D. (1990). Preference semantics: A family history. Report MCCS-90-194, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico.

Wilks Y. A., Fass D., Guo C.-M., McDonal J. E. , Plate T. and Slator B;M. (1993). Machine tractable dictionary tools. In Pustejovsky J. editors, *Semantics anc the Lexicon*. Dordrecht: Kluwer.