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 unit. 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.

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 ; Guthrie *et al.*, 1991 ; Wilks *et al.*, 1990). Véronis et Ide (1990) ont prolongé la méthode de Lesk en créant un réseau de neurones à partir des définitions du Collins English Dictionary. 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. 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 cooccurrences. 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 par 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.

À 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.

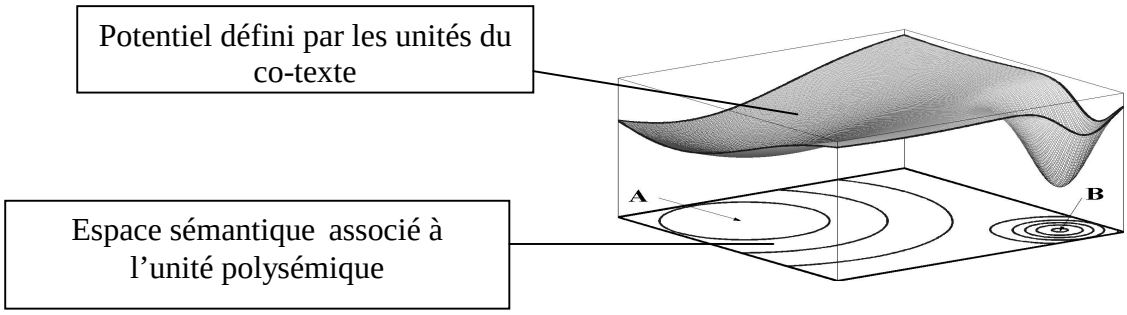

Figure 1. Représentation d'une fonction potentielle sur un espace sémantique bidimensionnel

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*.

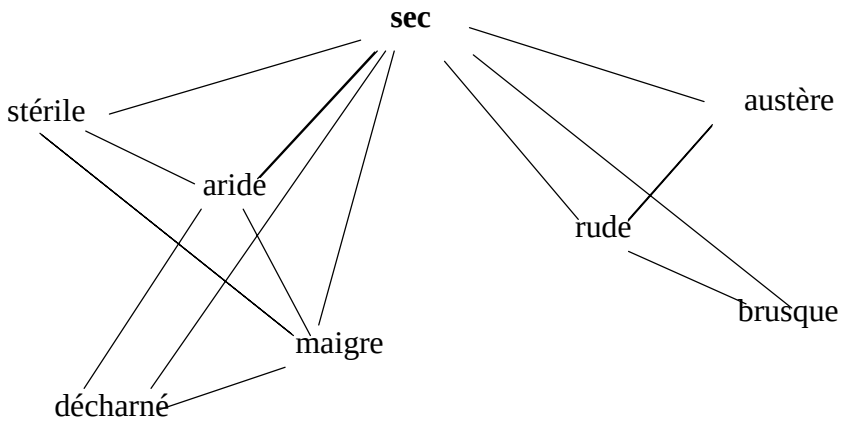

Figure 2. Un extrait du graphe de synonymie de sec

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.

4.2. *Zone de pertinence associée à un synonyme.*

À 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. Elle est calculée sur l'ensemble des cliques en donnant un poids égal à 1 aux cliques contenant ce synonyme, et un poids égal à –0.1 aux cliques ne le contenant pas.

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

$$f_j(x,y) = 1 - \max\left(0.1 - \sum_{k=1}^{c} a(k,j)\, e^{\frac{(x_k-x)^2-(y_k-y)^2}{\delta^2}}\right)$$ où (x_i,yi) sont les coordonnées du point

représentant la clique c_i dans l'espace sémantique, a(k,j)=1 si u_j appartient à la clique c_k, -0.1 sinon et $\delta = \frac{\max(dx, xy)}{10}$ où dx et dy mesurent les longueurs des intervalles de coordonnées.

À titre d'exemple, on trouvera en figure 4 et 5 les fonctions potentielles des adjectifs *brusque* et *

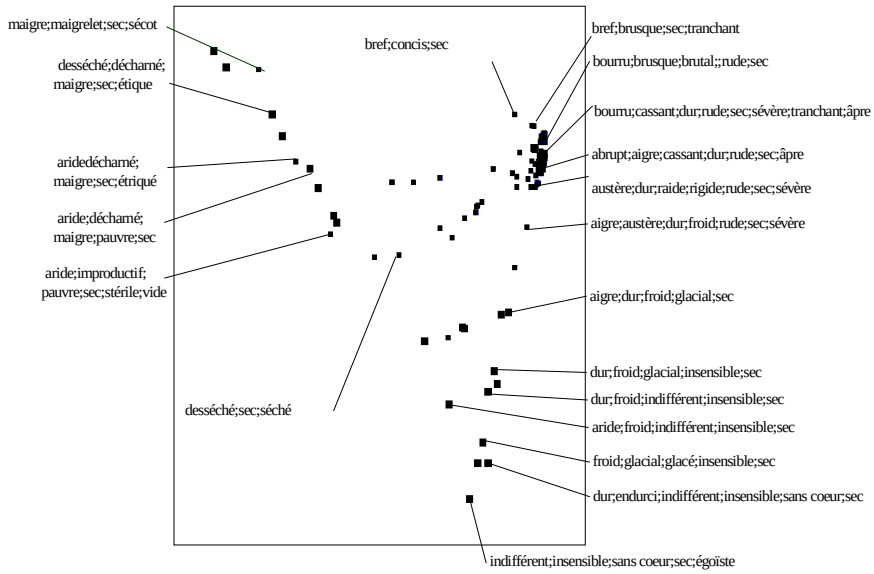

Figure 3. Espace sémantique de l'adjectif sec (63 synonymes, 94 cliques)

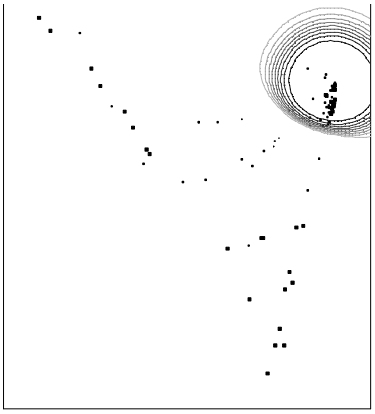

Figure 4.
Fonction potentielle associée à brusque

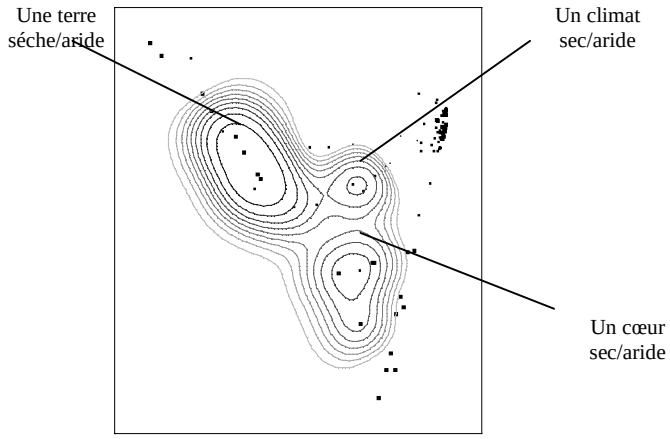

Figure 5.
Fonction potentielle associée à aride

On appelle n_{ij} le nombre réel d'occurrences du couple (e_i,uj) dans le corpus. S'il n'y avait pas d'affinité particulière entre certains noms et certains adjectifs, les couples seraient équidistribués. Le nombre d'occurrences de chaque couple (e_i,uj) ne dépendrait donc que de la fréquence des deux mots pris indépendamment dans le corpus. Appelons m_{ij} ce nombre moyen « théorique ». Alors on a : $m_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$ avec $n_{i\bullet} = \sum_{k=1}^{n} n_{ik}$ et $n_{\bullet j} = \sum_{k=1}^{m} n_{kj}$. Pour mesurer

l'affinité d'un nom et d'un adjectif, il faut donc comparer n_{ij} et m_{ij}. Nous définissons ainsi le degré d'affinité d_{ij} du nom e_i avec l'adjectif u_j : $d_{ij} = f(\frac{m_{ij}}{n_{ij}})$ où $f(x) = \frac{x}{2}$ ssi 0<x<2 et $f(x) = 1$ ssi x>2. Le degré d'affinité a_{ik} du nom e_i avec la clique c_k est donné par la formule suivante : $a_{ik} = \frac{\sum_{j=1}^{n} d_{ij} p_{ij} x_{kj}}{\sum_{j=1}^{n} p_{ij} x_{kj}}$ avec $p_{ij} = \frac{m_{ij}}{\sum_{k=1}^{c} x_{kj}}$ et $x_{kj} = 1$ ssi $u_j \in c_k$.

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.*

et les degrés d'affinité avec deux cliques contenant ces adjectifs : Sec ; bref ; brusque ; tranchant : 90 % et sec ; maigre ; maigrelet ; sécot : 12 %.

On 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. La valeur de la fonction potentielle associée au nom e_i au point de coordonnées (x ;y) est donnée par le calcul suivant :

$$g_i(x,y) = 1 - \max\left(0, \sum_{k=1}^{c} b(i,k)\, e^{\frac{(x_k - x)^2 + (y_k - y)^2}{\delta^2}}\right) \text{ où } b(i,k) = 2a_{ik} - 0.8 .$$

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.

Fleur sèche Ton sec

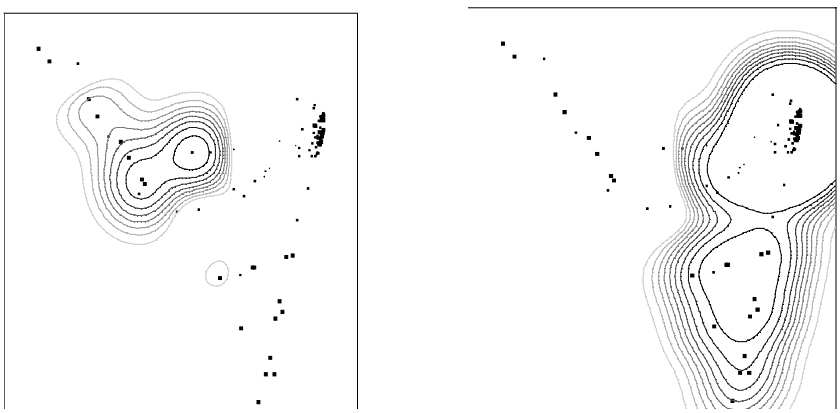

Figure 6. Potentiels désambiguisateurs des noms fleur et ton

4.3. *Évaluation*

Nous avons voulu évaluer la pertinence de nos calculs en comparant les résultats obtenus automatiquement à ceux obtenus par des locuteurs du français. Nous avons donc conçu une tâche de désambiguïsation réalisable à la fois par Visusyn et par des sujets. Il s'agit de sélectionner parmi les 5 synonymes proposés, celui ou ceux qui décrit le mieux le sens de *sec* en présence d'un nom donné. Les synonymes ont été sélectionnés en fonction de leurs indices de similitude. Ces indices sont calculés à partir du graphe de synonymie. L'indice de similitude entre deux mots est égal au nombre de liens qu'ils ont en commun, divisé par le nombre de liens qu'ils ont à eux deux. Il indique donc leur proximité sémantique. On calcule deux indices de similitude entre un synonyme de *sec* et *sec* lui même. Le premier i_1 est obtenu en comptant les liens avec l'ensemble des synonymes de *sec*, l'autre i_2 en restreignant à l'ensemble des synonymes du synonyme considéré. Pour notre travail, nous avions besoin de synonymes de *sec* qui ne soient pas trop ambigus. C'est pourquoi nous avons choisi des synonymes dont la polysémie est plus faible que celle de *sec*, c'est à dire des synonymes pour lesquels i_2 est supérieur à i_1. D'autre part nous avons veillé à ce que ces synonymes correspondent à des cliques bien réparties sur l'espace sémantique. Nous avons ainsi sélectionné les cinq adjectifs : *décharné, desséché, brusque, glacial, stérile.*

Pour réaliser la tâche, Visusyn calcule un taux d'adéquation entre le nom et l'adjectif en mesurant le recouvrement des fonctions potentielles du nom et de l'adjectif. Ce taux est ensuite arrondi et permet d'attribuer 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. Les sujets doivent eux aussi évaluer l'adéquation entre le synonyme proposé et le nom considéré. Ils attribuent à chaque synonyme une note entre 1 et 5 selon que le fait de remplacer *sec* par ce synonyme dans le contexte du nom considéré provoque un changement de sens important ou pas. On attribue ensuite un taux de réussite permettant d'évaluer la performance de Visusyn. On considère que 1 et 2 sont des notes proches puisqu'elles signifient que l'adjectif considéré peut être utilisé comme synonyme, plus ou moins adéquat, de *sec* dans le contexte du nom considéré. De même 3 et 4 sont liées et signifient que l'adjectif est rejeté plus ou moins vivement. En revanche 3 et 2 sont faiblement liées par la notion de changement de sens. Le tableau ci dessous récapitule les règles d'attribution du taux de réussite.

Note sujet	1				2				3				4			
Note Visusyn	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
% de réussite	100	80	0	0	80	100	20	0	0	20	100	80	0	0	80	100

Une partie des résultats est présentée ci dessous. On a indiqué pour chaque adjectif l'arrondi de la moyenne des notes données par les sujets :

nom	**adjectif**	**note Visusyn**	**note sujets**	**réussite (%)**		**nom**	**adjectif**	**note visusyn**	**note sujets**	**réussite (%)**
coup	brusque	1	1	100		**terre**	brusque	4	4	100
	décharné	4	4	100			décharné	3	3	100
	desséché	4	4	100			desséché	1	1	100
	stérile	4	4	100			stérile	2	2	100
	glacial	4	4	100			glacial	3	4	80
			moyenne	**100**					**moyenne**	**96**

nom	adjectif	note Visusyn	note sujets	réussite (%)		nom	adjectif	note visusyn	note sujets	réussite (%)
fleur	brusque	4	4	100		**boue**	brusque	4	4	100
	décharné	4	3	80			décharné	4	4	100
	desséché	3	1	0			desséché	4	1	0
	stérile	2	4	0			stérile	4	4	100
	glacial	4	4	100			glacial	1	4	0
			moyenne	**56**					**moyenne**	**60**
manière	brusque	1	2	80		**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	2	0			glacial	3	4	80
			moyenne	**76**					**moyenne**	**72**

Le taux de réussite globale est de 79 %. 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 majorité des cas, nous pouvons déterminer s'il s'agit d'un synonyme parfait ou approximatif (attribution correcte de la note 1 ou 2) : *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. 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. C'est le cas de *manières glaciales* par exemple. 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. C'est pour cette raison que nous ne réussissons pas à calculer le sens *fleur desséchée* pourtant attesté par les sujets. 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. Nous avons rencontré le problème une première fois dans le cas de *lit.* L'analyse de ce cas nous a conduit à chercher quelles étaient les cliques ayant un degré d'affinité avec *lit* supérieur à 90 %. Ce sont toutes des cliques contenant un des adjectifs *dur, rude* ou *froid*. Or *lit froid*, *dur,* ou *rude* ne sont pas synonymes de *lit sec*. On se heurte ici au problème de la polysémie de *lit*, « meuble pour dormir » ou « cours de rivière ». Or associer *sec* à *lit* oblige, du moins dans le corpus étudié, à choisir le sens « lit de rivière » et par la même à exclure pour *sec* les sens *froid, dur* et donc *brusque*. La question qui se pose ici est : comment notre système va t'il opérer cette double sélection ? Sans doute d'autres éléments du contexte entreront-ils en ligne de compte dans le calcul, mais au bout du compte il lui faudra bien décider qu'un « lit de rivière » ne peut être qu'*asséché*. D'où tiendra t'il ce savoir ? De quel type d'apprentissage ?

On trouve un problème un peu similaire avec *boue*. Visusyn considère 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 synonyme de *temps glacial*. Ici la polysémie du nom n'entre pas en jeu. Le problème se pose 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 avec *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 %. Nos principales voies de travail sont actuellement l'agrandissement du corpus et le réajustement des paramètres de calcul, notamment à partir de résultats d'expériences psycholinguistiques menées à plus grandes échelle par des experts. En outre 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. Enfin les problèmes rencontrés avec *lit* et *boue* nous pousse à approfondir notre réflexion théorique. Dans le cas de *lit*, il s'agit de trouver comment la polysémie du nom régissant entre en jeu. C'est tout l'aspect dynamique du calcul du sens qui apparaît ici, *sec* et *lit* s'influencent mutuellement. Ce qu'il s'agit de développer maintenant, c'est un autre outil informatique à base de réseaux connexionnistes récurrents, qui calcule de manière automatique les contraintes réciproques des unités polysémiques apparaissant dans un même énoncé. 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 présentant une structuration locale riche et une distance moyenne très petite sur l'ensemble du graphe ainsi qu'une structure hiérarchique) 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). Étude des critères de désambiguïsation sémantique automatique : résultats sur les cooccurrences. In *Actes de TALN 2003* : 35-44.

Bourigault D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de TALN 2002.*

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

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

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

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

Jacquet G. (2003). Polysémie verbale et construction syntaxique : étude sur le verbe *jouer*. In *Actes TALN 2003* : 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* : 101-106.

Inkpen D.Z. et Hirst G. (2003). Automatic sense disambiguation of the near-synonyms in a dictionary entry. In *Proceedings of the 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 *Proceedings of the 19986 SIGDOC Conférence*. Association for Computing Machinery *:* 24-26.

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

Ravin Y. et Leacock C. (2000). *Polysemy : Theoretical and Computational Approaches*. Oxford University Press**.**

Reymond D. (2001). Dictionnaires distributionnels et étiquetage lexical de corpus. In *Actes de RECITAL 2001 :* 24-33.

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

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

Victorri B. et Fuchs C. (1996). *La polysémie, construction dynamique du sens*. 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 Nirenburg S. (Ed.), *Machine Translation*. Cambridge University Press.

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

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