RECITAL 2004, Fès, 19- 22 avril 2004

Géométriser le sens.

Fabienne Venant

LaTTiCe – ENS
1 rue Maurice Arnoux 92120 Montrouge
fabienne.venant@ens.fr

Résumé – Abstract

Les recherches en sémantique lexicale s'appuient de plus en plus sur des ressources électroniques de grande taille (dictionnaires informatisés, corpus, ontologies) à partir desquelles on peut obtenir diverses relations sémantiques entre unités lexicales. Ces relations sont naturellement modélisées par des graphes. Bien qu'ils décrivent des phénomènes lexicaux très différents, ces graphes ont en commun des caractéristiques bien particulières. On dit qu'ils sont de type petit monde. Nous voulons mener une étude théorique mathématique et informatique de la structure de ces graphes pour le lexique. Il s'agit de les géométriser afin de faire apparaître l'organisation du lexique, qui est implicitement encodée dans leur structure. Les outils mis en place sont testés sur le graphe du dictionnaire électronique des synonymes (www.crisco.unicaen.fr). Ils constituent une extension du logiciel Visusyn développé par Ploux et Victorri (1998).

Research in lexical semantics tends to rely on large-scale electronic language resources (machine-readable dictionaries, corpora, ontologies), from which one can get varied semantic relationships between lexical units. Such relationships are naturally modelled by graphs. Although they describe different lexical phenomena, these graphs share some very specific characteristics. They are called "small worlds". We want to carry out a theoretical mathematical and informatic study of these graphs. The point is to geometrise them in order to reveal the organisation of the lexicon that is encoded in their structure. The tools we developed are tested on the graph of the electronic dictionary of synonyms (www.crisco.unicaen.fr). They represent an extension of Visusyn, the software developed by Ploux and Victorri (1998).

Mots-clefs – Keywords

Lexique, espace sémantique, synonymie, graphes petit monde.

Lexicon, Semantic space, synonymy, graph, small world.

1 Les graphes lexicaux.

Les travaux que nous avons menés récemment en désambiguïsation automatique (Victorri et al., 2003, Jacquet, 2003 ; Venant, 2004) nous ont amenés à nous intéresser de près aux graphes lexicaux. Grâce au développement de nouvelles technologies informatiques, les recherches en traitement automatique des langues s'appuient de plus en plus sur des ressources lexicales à grande échelle (corpus, ontologies, dictionnaires électroniques …). Ces ressources permettent d'obtenir de façon automatique des informations sémantiques sur les mots et les relations qu'ils entretiennent entre eux. Ces relations peuvent être représentées naturellement par des réseaux lexicaux. Les sommets en sont les mots d'une langue. Il existe plusieurs types de réseaux selon la relation lexicale utilisée pour définir les arcs du réseau. 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 (Véronis, 2003). Elle peut être de type paradigmatique comme c'est le cas dans le graphe sur lequel nous travaillons (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).

Aussi divers soient-ils, ces graphes partagent entre eux, et avec tous les autres graphes "de terrain" (réseaux sociaux, Internet, Web, réseaux électriques, réseaux de neurones,…) une structure et une topologie très particulières. On les appelle des graphes "petit monde". La théorie des graphes s'est pour l'instant très peu souciée de ces grands graphes (ils peuvent avoir plusieurs milliers de sommets ce qui est énorme comparé aux graphes habituellement étudiés en informatique théorique). Or nous pensons que la structure particulière de ces graphes est porteuse d'une information très riche sur les phénomènes sous jacents. Avoir accès à la structure d'un graphe lexical permettrait non seulement d'avoir une meilleure connaissance de l'organisation du lexique mais aussi d'automatiser l'accès à cette connaissance, ce qui peut être fondamental pour des systèmes de désambiguïsation automatique comme le nôtre. C'est pourquoi nous voulons « géométriser » ces graphes, c'est à dire les plonger dans un espace bi ou tri dimensionnel qui rende compte de leur topologie. Nos outils sont mis au point sur le graphe de synonymie fournie par le CRISCO (www.unicaen.crisco.fr) qui est au centre de nos travaux en modélisation de la polysémie et calcul dynamique du sens.

2 Les graphes de terrain

2.1 Graphes petit monde

Les graphes traditionnellement étudiés sont soit complètement réguliers soit complètement aléatoires. Dans un graphe régulier, chaque sommet a le même nombre d'arcs qui joignent un petit nombre de voisins dans un motif très clusterisé. Dans un graphe aléatoire chaque sommet est connecté arbitrairement à des sommets qui eux-mêmes se connectent aléatoirement à d'autres sommets. L'introduction des graphes aléatoires par Paul Erdös a permis de faire considérablement avancer l'étude des grands graphes (graphes présentant plusieurs milliers de sommets). Cependant il reste très insatisfaisant de modéliser un réseau réel par un graphe aléatoire. En fait la plupart des réseaux réels sont intermédiaires entre les réseaux ordonnés et les réseaux aléatoires. C'est pourquoi Watts et Strogatz (1998) ont

cherché un modèle qui leur corresponde mieux. Ils ont ainsi défini ce qu'on appelle les «petits mondes » et ont déterminé des paramètres permettant de les caractériser. Le concept de petit monde formalise le fait que même quand deux personnes n'ont aucun ami en commun, il n'y a qu'une petite chaîne d'amis qui les séparent. Ramené aux graphes ce résultat se traduit par le fait que la distance entre deux sommets quelconque est faible en moyenne. Ce phénomène est surprenant mais non caractéristique d'une organisation. Erdös et Reniyi (1960) ont en effet montré qu'on le trouve dans les graphes aléatoires. Il fallait donc pousser un peu plus avant pour caractériser les graphes de terrain. Ce qui est étonnant donc ce n'est pas tant que le monde est petit mais qu'il le soit bien que chacun d'entre nous possède un groupe de connaissances très resserré, dont la taille est faible par rapport à la population totale, et au sein duquel les gens ont de forte chance de se connaître entre eux. Formellement, cela se traduit par le fait que dans le graphe correspondant si A est relié à B et B est relié à C alors A a plus de chance d'être relié à C qu'à n'importe quel autre sommet du graphe. C'est ce qu'on appelle le clustering. Les graphes aléatoires sont faiblement clusterisés. Les graphes réguliers le sont fortement.
Ce qui va caractériser nos graphes de terrain, et en faire quelque chose d'intermédiaire entre les graphes réguliers et les graphes aléatoires, c'est qu'ils sont peu denses et possèdent à la fois une distance moyenne courte et un fort taux de clustering. C'est pourquoi Watts et Strogatz ont choisi pour caractériser les « petits mondes » les deux paramètres L et C :

- L, distance moyenne entre deux sommets, est un indice de la connectivité globale : L est donc très grand pour un graphe régulier et très petit pour un graphe aléatoire.

- C, coefficient de clustering, est un indice de la richesse de la cohésion locale. Il est défini de la manière suivante : si un sommet S a k voisins alors il peut exister au maximum n= k(k-1)/2 arcs entre ces k sommets. Soit m le nombre d'arcs qu'il y a effectivement entre ces k sommets alors le coefficient de clustering C_S associé au sommet S est m/n. Le coefficient global C est à égal à la moyenne des C_S quand S parcourt l'ensemble des sommets du graphe.

Pour savoir si on a affaire à un graphe de type petit monde, on compare les coefficients C et L à ceux d'un graphe aléatoire ayant le même nombre de sommets (n) et le même nombre moyen d'arcs par sommets (k). Pour un graphe petit monde on a $C >> C_{aléatoire} \cong k/n$ alors que L est du même ordre de grandeur que $L_{aléatoire} \cong \ln(n)/\ln(k)$

2.2 Graphes sans échelle.

Les travaux de Watts et Strogatz ont attiré l'attention sur les graphes de terrain. On a cherché à mieux les caractériser encore. Babarabasi et al. (1999) ont ainsi montré qu'ils font partie d'une autre classe très intéressante de graphes, les graphes sans échelle. Cela signifie que la répartition des degrés des sommets suit une loi de puissance : la probabilité P(k) qu'un sommet du graphe considéré aie k voisins décroît en suivant une loi de puissance $P(k)=k^{-\lambda}$ où λ est une constante caractéristique du graphe, alors que dans le cas des graphes aléatoires, c'est une loi de Poisson qui est à l'œuvre. La structure sans échelle se traduit donc par la présence d'un très grand nombre de sommets de faible degré et d'un nombre faible mais non négligeable de sommets de très haut degré. Ceci donne aux graphes sans échelle une structure qui peut être vue comme *hiérarchique* : localement, des sommets de très haut degré sont reliés à des sommets de moins haut degré, eux-mêmes reliés à des sommets de degré encore moindre, et ainsi de suite jusqu'à la masse des sommets de très faible degré. Les lois de puissance sont depuis considérées par de nombreux analystes de graphes comme la signature

de l'activité humaine. Ces premiers travaux ont suscité l'enthousiasme des théoriciens et beaucoup d'études ont été menées qui analysent des graphes divers des sciences sociales ou de la biologie. Cependant Gaume (2003) est le premier à mettre en évidence la structure de petit monde hiérarchique des graphes lexicaux. L'idée qui sous tend ses travaux est d'exploiter cette structure pour accéder de manière complètement automatique à une meilleure connaissance de l'organisation du lexique. C'est dans le même esprit que nous travaillons.

3 Le petit monde de la synonymie

3.1 La construction dynamique du sens.

Le cadre théorique de nos travaux est celui de la construction dynamique du sens défini par Victorri et Fuchs (96). Ce modèle, destiné à rendre compte de la place centrale de la polysémie dans la construction du sens, associe à chaque unité polysémique un espace sémantique dans lequel sont organisés ses différents sens. Les unités du cotexte induisent une dynamique sur cet espace. Ce sont les bassins de la fonction potentielle définie par cette dynamique qui permettent de définir la zone de l'espace sémantique correspondant au sens pris dans l'énoncé par l'unité considérée. Selon la largeur et le nombre de bassins on peut ainsi obtenir un sens précis, une ambiguïté ou une indétermination.

3.2 Le graphe des synonymes.

Ploux et Victorri ont mis au point Visusyn un logiciel permettant de construire de façon totalement automatique l'espace sémantique correspondant à un mot polysémique donné. Ce logiciel repose sur l'analyse du graphe du dictionnaire électronique des synonymes (D.E.S.) du laboratoire CRISCO. La base de départ est constituée de sept dictionnaires classiques (Bailly, Benac, Du Chazaud, Guizot, Lafaye, Larousse et Robert) dont ont été extraites les relations synonymiques. Les sommets du graphe sont des mots de la langue française. Deux mots sont reliés par un arc lorsqu'un des dictionnaires signale une relation synonymique entre eux. Le graphe correspondant possède 49133 sommets et 198549 arcs. Il s'agit bien d'un graphe peu dense, c'est à dire qu'il a relativement peu d'arcs relativement au nombre de ses sommets. Son degré moyen est 8.1. Le calcul des indicateurs L et C de Watts et Strogatz le classe dans la catégorie des petits mondes. On a en effet L= 4.7306 (qui est bien du même ordre de grandeur du L d'un graphe aléatoire $L_{al} = \frac{\ln(8.1)}{\ln(49133)} \approx 5.17$) et C= 0.35 (ce qui est très supérieur à ce qu'on aurait pour un graphe aléatoire c'est à dire $C_{al} \approx \frac{8.1}{49133} \approx 1.6 \times 10^{-4}$).

Enfin nous avons vérifié que la distribution des degrés suit une loi de puissance. C'est donc cette structure de graphe petit monde sans échelle qu'il va nous falloir exploiter dans la mise en place de nos outils de visualisation. L'objectif est double puisque l'algorithmique des petits mondes en est encore à ses prémices. Nos outils pourraient dépasser le cadre du lexique et s'appliquer à d'autres graphes des sciences humaines, pour peu qu'ils soient eux aussi des graphes petit monde sans échelle.

3.3 Premières visualisations obtenues

Pour déterminer automatiquement les paramètres de l'espace sémantique associé à une unité polysémique, Visusyn analyse le sous graphe dont les sommets sont l'unité étudiée et tous ses synonymes. Le principe sera illustré sur l'adjectif *sec*. L'idée est que ce sous graphe contient dans sa structure toute la sémantique de ce mot. Le travail consiste alors à définir la méthode

de géométrisation qui va faire apparaître cette structure dans une représentation en deux dimensions, la difficulté étant de trouver l'outil de la théorie des graphes qui va être pertinent. Ploux et Victorri ont eu l'idée d'utiliser les cliques. Une clique est un sous graphe complet maximal, c'est à dire un ensemble de sommets, le plus grand possible, reliés deux à deux. Chaque clique correspond à une nuance possible de sens pour *sec*. On va donc rechercher toutes les cliques du sous graphe. Notre espace sémantique est une projection en deux dimensions du nuage formé par ces cliques dans l'espace multidimensionnel engendré par les synonymes de l'unité lexicale considérée (pour les détails techniques consulter Ploux et Victorri, 1998). On trouvera une représentation de l'espace sémantique de *sec* en figure 1.

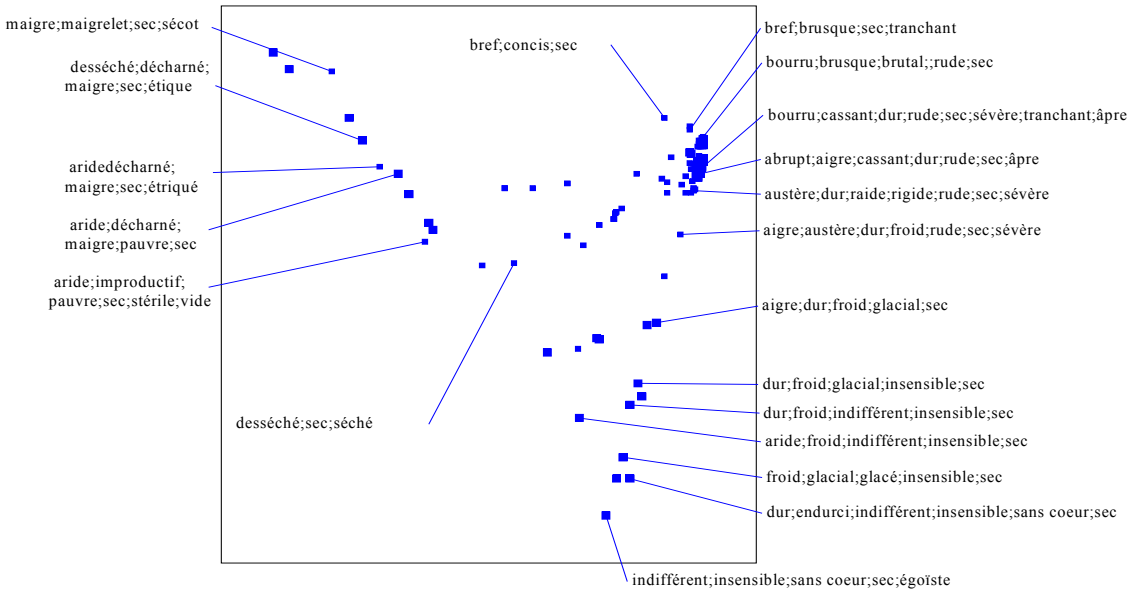

Figure 1: espace sémantique de sec (63 synonymes, 94 cliques)

La sémantique de *sec* étant bien connue, nous pouvons ici valider notre hypothèse de travail en vérifiant qu'on a bien obtenu, de façon totalement automatique, à partir de la topologie du graphe, une visualisation qui rende compte des différents sens de *sec* et qui les organise en fonction de leur proximité sémantique. On peut en effet regrouper les sens de *sec* en six acceptions principales, que l'on retrouve sur notre figure.

1. qui manque d'eau : *du sable sec* (centre de l'espace)
2. maigre, décharné : *un homme grand et sec* (en haut du quart supérieur gauche)
3. stérile, improductif : *rester sec aux questions du professeur* (en bas du quart supérieur gauche)
4. qui manque de sensibilité, qui ne se laisse pas attendrir, égoïste : *un cœur sec* (quart inférieur droit et bas du quart supérieur droit)
5. bref, abrupt, qui manque de douceur : *un coup sec* (haut du quart supérieur gauche)
6. seul : *un atout sec* (centre haut)

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. L'organisation des différentes cliques au sein de l'espace sémantique rend compte de ces relations.

3.4 Du local au global.

Le problème des visualisations obtenues est qu'elles sont locales. On ne peut visualiser le graphe de synonymie qu'au voisinage d'un de ses sommets. Le système de désambiguïsation sur lequel nous travaillons est une extension de Visusyn. Pour désambiguïser un mot il s'appuie d'une part sur l'espace sémantique calculé par Visusyn et d'autre part sur des calculs de cooccurrences issues de la base Frantext catégorisée. Pour améliorer ses performances, nous aimerions pouvoir lui fournir des informations plus globales rendant compte des subtilités dans les relations de synonymie. Il a par exemple rencontré le problème suivant: la tâche consistait à juger si un synonyme de *sec* était valable dans le contexte d'un nom donné (cf. Venant, 2004). Se basant uniquement sur des indices contextuels notre système, ayant rencontré dans le corpus beaucoup de *boue sèche* et beaucoup de *boue glaciale,* a considéré que *glacial* était un bon synonyme de *sec* dans le contexte de *boue*. (Notons que l'on rencontre le même phénomène avec *temps* : *temps sec* n'est pas synonyme 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. Face à ce genre d'erreurs, il nous semble crucial d'avoir des informations plus globales sur le graphe. Si on pouvait par exemple visualiser la "carte sémantique" des relations entre *sec*, *froid*, *glacial* et *glacé* et rendre les informations qu'elle contient accessibles automatiquement par des calculs de fonctions potentielles telles que celles que nous utilisons déjà, nous pourrions considérablement améliorer les performances de notre système. Nous cherchons donc actuellement à définir des méthodes permettant de visualiser des parties plus importantes du graphe. En utilisant des structures plus « lâches » que les cliques, les «noyaux», on doit pouvoir obtenir des cartes du graphe à différentes échelles, de la plus locale (celle que nous obtenons actuellement) à la plus globale. Nous voudrions construire ainsi une sorte d'atlas sémantique du français. L'idée est de dégager à partir des cartes globales des dimensions sémantiques organisatrices du lexique et de les utiliser pour structurer les cartes plus locales.

4 Géométriser le sens.

L'archétype du graphe petit monde, celui pour lequel on visualise le mieux la structure qui est en jeu, est le réseau social. Les sommets en sont des personnes. Elles sont reliées lorsqu'elles sont en relation soit amicale soit professionnelle, bref quand elles se rencontrent. On comprend très bien le côté à la fois hiérarchisé, peu dense globalement et très clusterisé localement. Chaque personne a en effet son cercle de connaissances, très clusterisé donc. Une d'entre elles est libre de rejoindre un club ou de déménager dans une autre ville. Il se forme alors de nouvelles connexions, des « raccourcis » entre diverses zones denses du graphe. Ces relations humaines sont cependant contraintes par un espace géographique constitué de villes plus ou moins importantes et plus ou moins éloignées les unes des autres. Plus une ville est importante, plus les activité de travail sont denses dans ce lieu. Plus deux villes sont proches, plus il y a d'activités qui impliquent les deux villes à la fois. Enfin chaque personne qui travaille se déplace en fonction de sa ou ses activités (une personne peut avoir plusieurs activités, changer d'emploi, etc.). Certains sont amenés à couvrir tout le territoire sur lequel s'exercent leurs activités, alors que d'autres sont cantonnés dans une partie seulement de ce territoire. Une rencontre est un événement qui se produit chaque fois que deux personnes se retrouvent dans le même lieu dans le cadre de leurs activités.
Notre hypothèse est qu'on peut de même définir un espace sémantique sous jacent aux rencontres lexicales. On peut considérer que les mots se rencontrent sur leur terrain d'activité, la parole. Les activités langagières s'y exercent chacune sur une région de sens, plus ou moins vaste suivant le sujet. Elles portent sur tous les domaines de l'expérience humaine, sur tous les sujets sur lesquels on communique par la parole Plus des sens sont voisins, plus ils ont de chance de correspondre aux mêmes activités langagières. Les mots sont au service des activités langagières. Certains sont utilisés dans beaucoup d'entre elles, d'autres dans une ou deux activités très précises. Certains couvrent toute la zone d'activité, d'autres seulement une partie. Pour que deux mots soient liés par la relation de synonymie partielle, il faut une probabilité suffisamment grande qu'ils se rencontrent dans un coin de l'espace sémantique.
Géométriser le sens c'est donc reconstruire l'espace sémantique à partir de la donnée des rencontres entre mots, c'est à dire à partir du graphe de synonymie. On veut pouvoir mettre en évidence « les villes », c'est à dire les zones denses en activités, et donc riches en rencontres. Ce que nous savons faire pour l'instant c'est repérer des zones d'activité très précises. Avec les cliques nous cherchons des groupes de mots qui se rencontrent deux à deux. Si l'on suit l'analogie, il y a dans l'espace sémantique des grosses villes et des villes moins importantes. Qu'est-ce qu'une grosse ville ? C'est une petite région de forte densité d'occupation, c'est-à-dire une région où ont lieu beaucoup d'activités, et donc où beaucoup d'individus se rencontrent. Bien sûr ils ne se rencontrent pas tous deux à deux : ils ne forment pas une seule clique. Mais le nombre de connexions entre eux doit être très grand, en tout cas plus grand qu'avec les autres individus. Si l'on veut visualiser l'espace à une grande échelle, on ne veut voir que les plus grosses villes et les distances entre elles. Autrement dit, ce ne sont plus les cliques qui nous intéressent, mais les grands sous graphes fortement connectés qu'on va appeler les d-noyaux (d'après un article de Guénoche sur la recherche de zone denses dans un graphe de protéines (Colombo et al., 2003)).

5 Les d-noyaux

5.1 Définition

Soit un graphe connexe G. On veut repérer dans G des ensembles de sommets fortement connectés mais pas forcément 2 à 2. On commence par définir une densité pour chaque sommet du graphe. Soit S un sommet, on considère le sous graphe formé par S et ses voisins immédiats. On note n_S le nombre de sommets de ce sous graphe et p_S son nombre d'arcs. La densité de S (nombre d'arcs par rapport au nombre d'arcs maximum possible) est donné par $d_S = 2\ p_S /(n_S .(n_S-1))$.

On appelle d-noyau tout sous graphe connexe de G tel que:

- tous ses sommets sont de densité $d_S \geq d$
- il est maximal pour la relation d'inclusion

On appelle d-noyau étendu toute extension connexe d'un d-noyau dont le degré moyen est supérieur ou égal au degré moyen du d-noyau dont il est l'extension. Quand d décroît, les d-noyaux étendus croissent en taille, passant de groupe très connectés au graphe tout entier.

Ce qui est intéressant, c'est qu'on peut appliquer les principes de visualisation des cliques aux d-noyaux étendus. Ainsi, à n'importe quelle échelle, on peut fabriquer un espace sémantique en associant à chaque d-noyau étendu un point de l'espace. La taille des noyaux pour un d donné est une indication précieuse : un grand d-noyau (relativement à la taille moyenne des noyaux pour le même d) peut être associé à un gros point, et on peut décider de ne visualiser que les plus gros points. Si tout se passe bien, on aura donc bien rempli l'objectif donné par l'analogie géographique : à une grande échelle, on ne voit que les plus grosses villes, et en diminuant l'échelle on voit plus de détails : les grosses villes deviennent des groupes de petites villes, d'autres villes apparaissent entre deux grosses villes.

5.2 Premiers résultats

Nous avons calculé des d-noyaux étendus (extension limitée aux voisins immédiats du noyau) pour d variant de 1 à 0.3 sur un sous graphe du dictionnaire formé de la manière suivante : on sélectionne d'abord le verbe du dictionnaire ayant le plus de synonymes (*faire*) ainsi que tous ses synonymes, parmi les verbes restants on sélectionne à nouveau celui qui a le plus de synonyme (*battre*) et tous ses synonymes, on réitère l'opération tant qu'on n'a pas atteint les mille sommets. La figure 2 montre la visualisation obtenue avec un seuil de densité égal à 1. La taille du rond représentant un noyau est proportionnelle au nombre de sommets constituant ce noyau. Pour l'étiquetage on ne fait apparaître que les sommets les plus caractéristiques, c'est à dire ceux qui ont le plus de liens vers l'intérieur du noyau et le moins possible vers l'extérieur. Les verbes s'organisent en un triangle dont on peut identifier les différents sommets. En bas à gauche se regroupent les verbes exprimant une action constructive : *produire* ou *construire* mais aussi *créer*, *exciter*, *attiser*. Beaucoup de ces verbes sont réunis dans un noyau assez important étiqueté par *faire* (ce noyau contient entre autres *donner*, *former*, *créer*, *façonner*, *former*…). A l'opposé le sommet du haut est résolument destructeur. Il s'agit d'ôter quelque chose : matière, valeur ou estime. On trouve ainsi : *évider*, *dévaluer*,

restreindre, *avilir*, *amoindrir*, *diminuer*. L'axe menant d'un pôle à l'autre passe par deux noyaux importants étiquetés par *battre* et *arrêter*. Le troisième sommet est consacré à la notion de départ : *décamper*, *partir*, *se sauver*…. Le passage de la notion dépréciative à la notion de départ se fait de façon subtile le long de l'arête du triangle : *d'amoindrir* à *s'affaiblir*, de *décliner* à *s'amenuiser* on finit par *disparaître*, *s'éclipser* et puis *s'enfuir*. On pourra noter que cette visualisation fait apparaître quelques noyaux importants qui résument l'organisation décrite ici. Ce sont *amoindrir-diminuer* ; *battre-rosser*, *arrêter*, *passer*, *faire*, *exhorter*, *décamper*. Lorsqu'on diminue le seuil de densité, le triangle se resserre petit à petit. Le rapprochement des sommets valorisants et dépréciatifs conduit à la formation d'un gros noyau étiqueté par *calmer-faire-moderer.* . Les capacités de calcul de nos machines ne nous ont pas permis d'aller au-delà d'un seuil de densité de 0.3. Le résultat est présenté en figure 3.

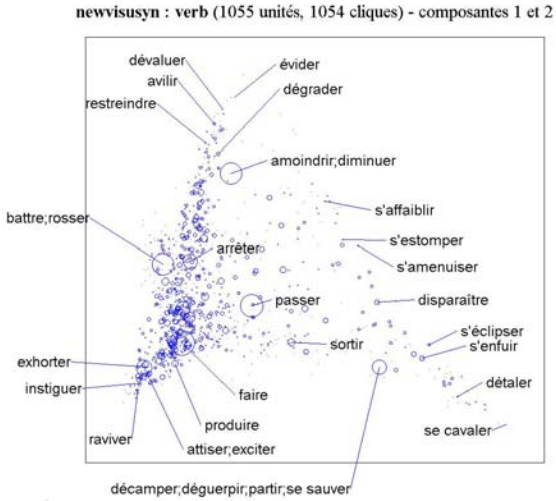

Figure 2: 1-noyaux obtenus pour 1055 verbes

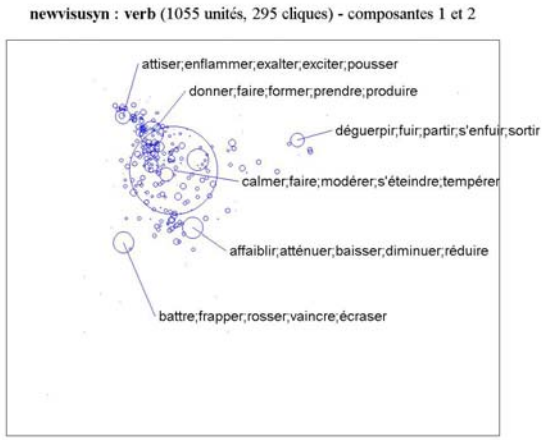

Figure 3: 0.3-noyaux obtenus pour 1055 verbes

5.3 Conclusions et perspectives.

Ces premières visualisations permettent déjà de faire émerger des dimensions sémantiques organisatrices du lexique verbal du français. On aurait pu imaginer en effet voir apparaître un axe opposant les verbes d'état aux verbes d'action ou encore que les dimensions aspectuelles seraient dominantes. Ce qui apparaît ici c'est que les verbes du français s'organisent selon des dimensions générales, de valuation ou de mouvement. Ce sont ces mêmes dimensions que l'on retrouve dans le triangle conceptuel des verbes du français que Gaume décrit dans ces derniers travaux (Gaume, 2004). Il l'a cependant obtenu suivant un mode opératoire radicalement différent du nôtre. Il transforme d'abord le graphe en une chaîne de Markov dont les états sont les sommets du graphe. Des particules se déplacent alors aléatoirement en empruntant les arcs du graphe et ce sont les dynamiques de leurs trajectoires qui donnent les propriétés structurelles du graphe. Le fait que nous ayons pourtant obtenu des résultats similaires nous donne à penser qu'il s'agit là d'une propriété forte de l'organisation des verbes du français. Elle est inscrite dans la structure du graphe et notre géométrisation permet de la mettre en évidence. Notre méthode en est encore à ses débuts mais le fait qu'elle fasse déjà apparaître des propriétés structurelles fortes nous laisse à penser qu'au fil de l'affinage elle va nous faire découvrir des propriétés de plus en plus subtiles. Une des questions que posent cependant ces premiers résultats est celle de la validation sémantique des représentations obtenues. Il nous faudra en particulier les valider sur des paradigmes lexicaux

sur lesquels nous disposons d'études menées par les linguistes de notre équipe. Une des pistes de validation envisagées est de les utiliser dans nos tâches de désambiguïsations, elles-mêmes validées à l'aide d'expériences psycholinguistiques. Les dimensions sémantiques dégagées sur la globalité du graphe, ou sur certains sous graphes importants (verbes, adjectifs,…), permettront d'orienter et d'affiner les représentations locales utilisées par notre système.

Il reste cependant un grand travail exploratoire à mener avant l'utilisation en TAL concernant notamment l'algorithme de calcul utilisé. Celui que nous avons présenté ici n'est sans doute pas le plus performant. Il peut être modifié de différentes façons (définition différente de la densité, critère d'arrêt dans la phase d'extension des noyaux…). Il nous faudra évaluer la pertinence de chacune des variantes aussi bien d'un point de vue formel, en s'interrogeant sur la nature des informations obtenues sur la structure du graphe, que d'un point de vue sémantique en évaluant en quoi les géométrisations obtenues sont valides. D'autre part l'algorithme actuel s'appuie essentiellement sur la nature clusterisée du graphe étudié. Nous savons que ce graphe possède d'autres caractéristiques importantes. Il faudra que l'algorithme final s'appuie davantage sur ces autres caractéristiques si nous voulons que nos outils soient effectivement utilisables sur les graphes de terrain en général. Enfin la notion de changement d'échelle est à travailler. Il faut sélectionner les noyaux à afficher en fonction de leur taille et mettre en place des systèmes de zoom qui puisse permettre de passer d'une visualisation à une autre.

References

BARABASI A-L, A R., JEONG H. (1999), Scale free characteristics of random networks: The topology of the World Wide Web. *Physica A*, 281:69.77, 2000.
BOURIGAULT D. (2002), Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. Actes *TALN 2002.*
COLOMBO T., GUENOCHE A, QUENTIN Y. (2003) . Recherche de zones denses dans un graphe. Application aux gènes orthologues. http://www.inist.fr/uir/jim03/colomb.pdf.
ERDÖS P. and RENIYI A. (1960), *Publ. Math. Inst. Hung.Acad. Sci 5*,17-61
GAME B. (2003), Analogie et Proxémie dans les réseaux petits mondes, *Regards croisés sur l'analogie*. RIA, n°spécial, Vol 5-6, Hermès Sciences.
GAUME B. (2004), Ballades aléatoires dans les Petits Mondes Lexicaux, *I3 Information Interaction Intelligence,* CEPADUES édition (à paraître).
GAUME B., DUVIGNAU K., GASQUET O. et GINESTE M-D. (2002). Forms of Meaning, Meanings of Forms. *Journal of Experiment and Theoretical Artificial Intelligence*, 14(1): 61-74.
Jacquet G. (2003). Polysémie verbale et construction syntaxique : étude sur le verbe jouer. Actes *TALN 2003*, pages 469-479.
MILGRAM S., (1967), The small world problem, *Psychol. Today 2*,60-67.
PLOUX S. et VICTORRI B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires informatisés des synonymes. *TAL*, 39(1) :161–182.
VENANT F. (2004). Polysémie et calcul du sens. Actes *JADT 2004* (à paraître).
VERONIS J. (2003). Cartographie lexicale pour la recherche d'information. Actes *TALN 2003*, pages 265-275.
VICTORRI B., FRANCOIS J., MANGUIN J.L (2003)., Dynamical construction of meaning in polysemic units.
VICTORRI B., FUCHS C. (1996), *La polysémie, construction dynamique du sens*, Paris, Hermès.
WATTS D.J., STROGATZ S.H. (1998), Collective dynamics of 'small-world' networks. *Nature* 393: 440-442 [1, 1998].