Une exploration de la structure sémantique du lexique adjectival français

Fabienne Venant

Une exploration de la structure sémantique du lexique adjectival français

Fabienne Venant

Laboratoires Lattice / Lalicc
ENS, 1 rue Maurice Arnoux 92120 Montrouge / 28 rue Serpente 75006 Paris

fabienne.venant@paris4.sorbonne.fr

RÉSUMÉ. *Cet article présente un travail en sémantique lexicale. L'idée est d'explorer un graphe lexical de façon à mettre en évidence la structure du lexique qu'il modélise, et de rendre ainsi l'information qu'il contient accessible à un système automatique. Le cadre théorique de l'exploration est géométrique, on construit un espace sémantique associé au graphe, la topologie de cet espace rendant compte de celle du graphe. Les outils sont développés et testés sur un graphe de synonymie adjectivale et s'appuient sur la structure petit-monde à invariance d'échelle de ce graphe. Ils sont destinés à être ensuite utilisés dans l'exploration d'autres graphes lexicaux et plus généralement d'autres graphes de terrain.*

ABSTRACT. *This paper presents an exploration of a lexical graph, within the framework of lexical semantics. The aim of the exploration is reveal the structure of the lexicon modelled by the graph so an automatic system can reach the information it contains. We developed geometric tools to build a semantic space associated with the graph. The topology of this space can account for the topology of the graph. These tools were developed using a graph of adjectival synonymy and its scale-free small-world structure. We now want now to use them to explore any lexical graph, and more generally any graph related to human activities.*

MOTS-CLÉS : *sémantique, lexique, adjectif, exploration de graphe, espace sémantique, petit-monde, invariance d'échelle*

KEYWORDS: *semantics, lexicon, adjective, graph exploration, semantic space, small-world, scale-free.*

1. Introduction

L'importance de la sémantique lexicale pour le TAL n'est plus à démontrer. Qu'il s'agisse de constitution de terminologies, de choix de collocations pour la génération de texte, ou de désambiguïsation automatique, les modélisations du lexique se sont multipliées au cours de ces dernières années. La question de la représentation du sens et de l'accès à l'information dans les ressources lexicales devient fondamentale pour la résolution de nombreux problèmes actuels. La sémantique lexicale en TAL doit se poser la double question du contenu - comment définir le sens lexical et quelle représentation en donner - et de l'organisation - comment prendre en compte les relations sémantiques entre les unités lexicales. Un des enjeux majeurs étant ensuite de pouvoir grâce à ces informations lexicales mettre en place une interface sémantique/syntaxe efficace et pertinente.

On a ainsi vu se développer depuis 1980 des ressources lexicales à grande échelle (corpus, ontologies, dictionnaires électroniques...). Ces ressources fournissent des informations sur les mots et les relations qu'ils entretiennent entre eux. Ces relations peuvent être de différents types. Sur l'axe paradigmatique par exemple, le Dictionnaire Electronique des Synonymes (DES : www.crisco.unicaen.fr) propose un éventail très complet des relations de synonymie (partielle) en français. En ce qui concerne l'anglais, la ressource de référence est WordNet (Felbaum, 1998 - http://wordnet.princeton.edu/), un système de références lexicales croisées. Les mots y sont organisés en ensembles de synonymes (Synsets) représentant le concept lexical sous-jacent. La relation sémantique de base dans WordNet est la synonymie, mais d'autres relations sémantiques comme spécifique/générique, hyperonyme/hyponyme ou encore la relation partie-tout sont utilisées pour structurer l'ensemble des Synsets. Les thesaurus (le plus célèbre étant le Roget's thesaurus, Roget, 1963 - http://thesaurus.reference.com/Roget-Alpha-Index.html) s'emploient aussi à rendre compte des relations entre mots. Un thesaurus est souvent associé à un domaine donné, et fournit une représentation hiérarchisée d'un vocabulaire spécialisé. Les dictionnaires électroniques, comme le *Trésor de la langue Française Informatisé* (TLFI - http://atilf.atilf.fr/tlf.htm) fournissent quant à eux des informations très complètes sur la sémantique des mots, et les relations qu'ils entretiennent. Le problème qui se pose alors est celui de l'accès aux données. Plus récemment, l'essor de la linguistique de corpus a permis la création de nouvelles ressources utilisant des relations syntagmatiques, avec l'étude des cooccurrences ou des collocations, ou distributionnelles, les mots sont rapprochés sur la base de contextes partagés. Une nouvelle problématique apparaît alors, celle de la constitution (automatique ou non) de catégories sémantiques, notamment en vue de créer des ontologies (mais d'autres applications sont possibles, en particulier dans le domaine de la désambiguïsation automatique). Citons enfin la base de données lexicales FrameNet (http://framenet.icsi.berkeley.edu/book/book.pdf) qui allie les principes de la sémantique des cadres (FrameSemantic) et de la lexicographie de corpus (British National Corpus). En utilisant à la fois des annotations manuelles et des résumés automatiques de ces annotations, elle fournit

pour chaque mot des informations sémantiques mais aussi syntaxiques (valence) sur chacun de ses sens possibles. De telles ressources, inscrites dans un cadre théorique précis, posent le problème de leur exploitation par des modèles se posant d'autres problèmes théoriques. Barque (2006) remarque ainsi à propos de FrameNet que « à l'issue de la description, les unités lexicales formant l'unité polysémique évoquent leur cadre respectif. FrameNet laisse donc de côté le délicat problème de l'organisation des sens d'une unité polysémique et n'explicite pas de manière systématique les liens de polysémie qui lient entre elles ces unités lexicales. »

On le voit à travers cet aperçu partiel, la création et l'utilisation d'une ressource lexicale est complexe et dépend beaucoup du type d'information qu'on y cherche ou de l'utilisation qu'on veut en faire. C'est pourquoi la plupart des modélisations actuelles sont des modélisation *ad hoc*. La taille des lexiques, la complexité des traitements, l'absence de consensus dans les représentations n'ont pas favorisé le développement de descriptions sémantiques du lexique en grandeur réelle (à part peut être WordNet et FrameNet). C'est que plus la quantité d'information présente est importante, plus l'accès à cet information est difficile. Or les outils TAL (notamment pour la Recherche d'Information) requièrent des connaissances linguistiques massives. Il est donc de plus en plus nécessaire de concevoir des modes de modélisation de ces données qui soient automatiques, ou qui puissent être assistés automatiquement, et de prévoir leur enrichissement continu par des mises à jour.

C'est dans ce cadre qu'on a vu se développer récemment l'utilisation de graphes. Le fait que le lexique soit essentiellement structuré par des relations (qu'elles soient sémantiques, morphologiques, syntaxiques ou grammaticales) induit naturellement une modélisation sous forme de réseau. Les sommets en sont les mots ou les concepts, les arêtes représentent les relations (synonymie, relation sémantique, contextes partagés…). La Figure 1 présente par exemple un extrait de graphe issu de WordNet.

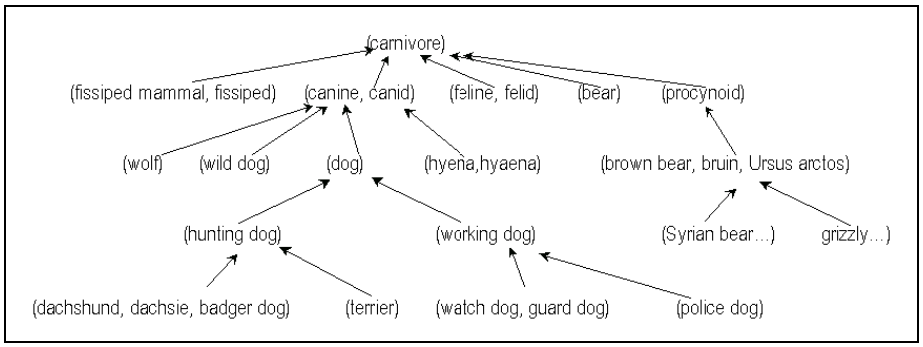

Figure 1. Un extrait de WordNet

Un exemple particulièrement représentatif et ingénieux de l'utilisation de WordNet est donné par les métriques heuristiques de "distance sémantique" entre les concepts d'une ontologie particulière, basées sur la distance à parcourir dans le graphe. Cette distance peut permettre de quantifier par exemple la similarité de deux concepts. Elle peut également servir à la désambiguïsation. Un autre exemple de graphe nous est donné par le logiciel FrameGrapher, associé à FrameNet. Il permet de visualiser les relations entre cadres, avec des « cadres pères » qui pointent sur des « cadres fils » . Véronis (2003) utilise quant à lui un graphe de cooccurrences pour rechercher des proximités sémantiques. Bouaud *et al* (2000) utilisent un graphe distributionnel comme « mise en forme du matériau textuel qui donne une vue d'ensemble du corpus. » Ce graphe est ensuite utilisé pour mettre en évidence le fonctionnement de certains mots en vue de faire ensuite émerger des catégories sémantiques. Gaume (2004) s'attelle à la lourde tâche de rendre accessible l'information riche contenue dans un dictionnaire. Il modélise le dictionnaire comme un réseau dont les sommets sont les entrées du dictionnaire, un lien existant entre deux mots quand l'un apparaît dans la définition de l'autre. L'analyse du graphe ainsi créé permet de rendre compte de la topologie complète du dictionnaire et non de quelques entrées ciblées.

L'intérêt d'une modélisation sous forme de réseau est double. D'une part il est facile d'insérer de nouvelles données dans un réseau, et donc de le mettre régulièrement à jour. D'autre part l'analyse du graphe sous-jacent ouvre de belles perspectives d'exploration et donc un accès facilement automatisable à l'information qu'il contient. L'objet du travail présenté ici est la mise au point d'une méthode d'exploration de grands graphes lexicaux. (Il reprend en partie le travail présenté dans Venant, 2006). Les différents graphes évoqués ci-dessus sont bien sûr de nature très différente, en fonction des relations qu'ils représentent. Mais aussi divers soient-ils, nombre d'entre eux partagent une structure et une topologie très particulières. On les appelle des graphes petit-monde. C'est sur cette structure que nous nous sommes appuyés pour développer notre méthode d'exploration. Nous avons développé et testé nos outils sur un graphe de synonymie, mais ils sont destinés à être utilisé sur n'importe quel graphe, pour peu qu'il soit de type petit-monde

2. Le lexique adjectival

Nous avons choisi de nous restreindre dans un premier temps au lexique adjectival du français. La mise au point des outils d'exploration demande en effet de travailler sur un lexique dont on connaît bien les caractéristiques, de façon à pouvoir analyser les visualisations obtenues. Le lexique adjectival constitue de ce point de vue un champ d'expérimentation idéal. De taille assez conséquente pour que son exploration automatique ait un sens, il est abondamment décrit dans la littérature, ce qui nous a permis de nous faire une première idée de sa structure. Les adjectifs constituent cependant une catégorie dont on n'est loin d'avoir élucidé tous les mystères. Si la littérature nous a permis de dégager les caractéristiques essentielles

de la sémantique adjectivale, nous espérons bien qu'en retour, notre exploration permettra de mieux comprendre la structure, très complexe, de cette catégorie.

2. 1 *Types sémantiques adjectivaux*

Le comportement adjectival est assez difficile à caractériser. Les frontières de cette catégorie sont à l'heure actuelle encore mal définies. L'adjectif a ainsi longtemps été assimilé à la catégorie des verbes (Platon) ou des noms (Aristote, Port Royal) même si on lui assignait une place à part dans cette catégorie (qualité *vs.* substance, possibilité de gradation). Depuis 1747 (Abbé Girard[1]), on le reconnaît comme une partie autonome du discours, mais les affinités qu'il possède avec les catégories des noms (existence d'adjectifs substantivés, de substantifs adjectivés), des verbes (participes passés et présents, valence des adjectifs, similarités distributionnelles) ou encore des déterminants rendent particulièrement floues les frontières de cette catégorie. C'est par l'étude du sémantisme que l'on va tenter d'y voir plus clair. Dès le milieu du XVIII[ème] siècle, le philosophe Du Marsais[1] s'est intéressé de près au sémantisme de l'adjectif. Dans son ouvrage *Logique et principe de grammaire* (1769), il affirme que la combinaison d'un substantif et d'un adjectif ne représente dans notre esprit qu'une seule idée : les deux parties essentielles en lesquelles elle se décompose sont conçues en même temps et font corps l'une avec l'autre. C'est le principe de la **qualification** qui est ainsi ébauché : l'unité dans la complexité. Reiner (1968) précise cette idée en posant que la signification du groupe nominal est le résultat de l'union interne de ses éléments constitutifs. C'est par l'accommodation sémantique mutuelle du substantif et de l'adjectif que le syntagme n'éveille dans l'esprit qu'une seule idée totale. Elle cite par ailleurs Walther Bame pour qui la signification d'un adjectif comprend à la fois des éléments affectifs et des éléments logiques. Pour lui, ce qui varie d'un adjectif à l'autre, c'est la qualité et la quantité des éléments de l'une ou de l'autre sorte, ce qu'il appelle la « puissance » de l'adjectif. Cette distinction sémantique est reprise par Marouzeau[1] (1922). Considérant les deux séries d'exemples suivants :

(1) *(le costume) féminin, (un animal) aquatique, (l'épopée) napoléonienne, (le territoire) français, (une montre) métallique, (une fleur) bleue,*

(2) (*un costume) étrange, (un) bel (animal), (une) magnifique (épopée), (un) riche (territoire),*

Il écrit : « dans la première série d'exemples la qualité appartenait en propre à l'objet, indépendamment de notre appréciation ; dans la seconde série, elle n'existe qu'en tant qu'elle est ressentie par nous. L'adjectif a dans le premier cas une valeur *objective, intellectuelle,* et *subjective, affective* dans le second. » Ce critère sémantico-psychologique a été repris ensuite pour distinguer les adjectifs dits **qualificatifs** (subjectifs) des adjectifs **déterminatifs** (objectifs). Les adjectifs déterminatifs désignent donc des propriétés comme des rapports de temps ou de lieu, ou des qualités physiques (couleur, goût...).

[1] Cité par Goes 1999

2.1.1. *Les adjectifs de relation*

On distingue au sein des adjectifs déterminatifs une classe encore plus particulière, celles des adjectifs **de relation.** Daille (2001) rappelle que « la tradition linguistique et grammaticale distingue deux grandes catégories parmi les adjectifs : les adjectifs qualificatifs comme *important*, et les adjectifs relationnels comme *laitier*. Les premiers ne peuvent pas avoir une interprétation actancielle à la différence des seconds : l'adjectif *laitier* au sein du syntagme nominal *production laitière* est argument du nom prédicatif production, ce qui n'est pas le cas pour l'adjectif *important* dans le syntagme *production importante*. Le terme d'adjectif de relation ou relationnel a été introduit par Bally (1965) et permet d'exprimer cette idée de « relation » habituellement exprimée par une préposition. Daille donne aussi l'exemple de *municipal* dans *parc municipal*. L'appartenance du parc aux parcs municipaux n'est pas due à une appréciation subjective, alors qu'une qualité comme *admirable* (*un parc admirable*) manifeste un point de vue subjectif. Fondamentalement, les adjectifs de relation relèvent de la détermination du nom, les adjectifs qualificatifs participant à la caractérisation du référent. Selon elle, c'est en relation avec le locuteur que l'on pourrait construire des objets auxquels réfèrerait l'auteur en fonction de ses jugements. Tout auteur de jugement s'expose à la polémique : poser que tel film est un film remarquable, à voir, peut provoquer la réplique c'est un film à éviter, raté ; la référence est commune mais la caractérisation du film est différente. Les adjectifs qualificatifs et relationnels partagent les propriétés d'accord en nombre et en genre avec le nom qu'ils accompagnent et la possibilité d'occuper la fonction d'épithète. En revanche, ces deux classes se différencient à l'aide de propriétés morphologiques, paraphrastiques, syntaxiques et sémantiques qui s'appliquent soit à l'adjectif seul soit au groupe nominal dans lequel il apparaît.

2.1.2. *Les adjectifs primaires*

Parmi les adjectifs **qualificatifs**, une classe d'adjectifs fait l'objet d'une attention toute particulière, la classe des adjectifs **primaires**. Ils sont définis par Pottier (1985) comme les adjectifs qui expriment « les propriétés fondamentales des êtres et des choses. » Les avis divergent sur les critères discriminant les adjectifs primaires des autres adjectifs (dérivés ? non dérivés ? monosyllabiques ? disyllabiques ? trisyllabiques ?) mais tous s'accordent pour y ranger les adjectifs exprimant des données immédiates des sens, et des dimensions sémantiques évaluables ou spécifiables. On retrouve, à travers les différentes études, les mêmes concepts exprimés par les adjectifs primaires : *grand, petit, long, court, nouveau, vieux, bon, mauvais, noir, blanc, rouge, cru/vert/non mûr.* Présents même dans les langues qui n'ont qu'une catégorie limitée d'adjectifs, ils constituent une catégorie « spéciale » dans les langues où les adjectifs sont nombreux. Le nombre d'adjectifs considérés comme primaires varie beaucoup selon les auteurs (Reiner 1968, Borodina 1963, Noailly 1999…). La constante dans toutes les listes établies est que les adjectifs sélectionnés correspondent à l'idée que l'on se fait généralement de ce que devrait être un adjectif. Si on demande en effet à des locuteurs du français de citer 5

adjectifs sans réfléchir, les plus fréquemment cités sont : *petit, grand, bon, mauvais, joli* (expérience faite par Goes, 1999, auprès de ses étudiants et par moi-même auprès de mes proches). Ils présentent la caractéristique de prendre parfois un sens très général et n'être plus que de simples intensifs au sens vague qui se ressemblent les uns les autres (*un* ***grand*** *lecteur* ≈ *un* ***énorme*** *lecteur*, ***haut*** *goût* ≈ ***bon*** *goût*, *deux* ***bonnes*** *heures* ≈ *deux* ***grandes*** *heures*). C'est ce que Goes appelle la désémantisation. De plus, ces adjectifs s'emploient majoritairement en position antéposée alors que la grande majorité des adjectifs du français préfèrent nettement la postposition.

2.1.3. *Les adjectifs intensifs*

Romero (2004) propose de dégager une troisième classe, les adjectifs **intensifs.** Il s'agit des adjectifs au moyen desquels on peut intensifier un nom :

- *énorme* (envie),
- (chaleur) *terrible*,
- (mystère) *insondable*,
- (beauté) *inénarrable*,
- (froid) *glacial*.

Les adjectifs intensifs sont eux aussi sujets au phénomène de désémantisation souligné par Goes. Ils sont la plupart du temps paraphrasables par *grand*. Quand celui-ci ne convient pas, c'est souvent pour des raisons stylistiques, et on peut alors utiliser *gros*, *fort* ou *vrai* et/ou une phrase contenant *très* (*Jacques a une énorme envie* = *Jacques a une très grande envie*). On les classe habituellement parmi les adjectifs qualificatifs. Mais en réalité, ce que fait l'adjectif intensif, c'est « une opération qui met en jeu la notion de degré (c'est-à-dire un cas particulier de quantification). Les adjectifs intensifs semblent donc échapper à la dichotomie traditionnelle qualificatif/de relation et ont un sémantisme bien particulier. Lorsqu'ils s'appliquent à un nom gradable, ou scalaire, leur action consiste à « situer l'occurrence en haut de l'échelle qui définit le nom » (*énorme envie* ne qualifie pas *envie*, ne le range pas non plus dans une classe, mais signifie qu'on se situe en haut de l'échelle des envies). Ils agissent aussi sur les noms *a priori* non gradables. Quand on parle *de vraies vacances,* on ne parle pas de la qualité des vacances (comme dans *vacances chères ou longues*) ni de tel ou tel type de vacances (d'hiver, touristiques) mais on dit que les propriétés qui constituent le sens de *vacances* sont réunies, ou intensifiées. L'intensivité n'exprime par forcément une gradation positive. Il existe des adjectifs « désintensifs »: *petite hausse, faible motivation, dimensions modestes*. Notons que *dérisoire* est intensif après *facilité*, désintensif après *salaire*.

2.2. *Type d'adjectifs ou type d'emplois ?*

On a parlé jusqu'ici d'adjectif qualificatif, relationnel ou intensif. En fait, il est souvent bien difficile de caractériser de façon définitive le comportement sémantique d'un adjectif. On aura, par exemple, du mal à trouver un adjectif purement intensif. Il y a bien sûr des adjectifs de nature intensive (ou statistiquement intensive), comme *extrême* (*extrême bonté*), mais qui présentent quand même des emplois qualificatifs (*expérience extrême* : qui comporte des risques ; *partie extrême :* qui se trouve au bout). *Méchant* est par ailleurs un adjectif qui par nature est qualificatif, mais on peut le trouver en emploi intensif dans *il vient de s'acheter une méchante voiture* ou désintensif dans *Il portait un méchant costume de laine.* Réciproquement un adjectif statistiquement relationnel comme *procédural* est utilisé de façon intensive dans *lenteur procédurale*. Par ailleurs, nombre d'adjectifs relationnels oscillent entre des emplois qualificatifs et des emplois relationnels. Prenons l'exemple de l'adjectif *populaire.* On le trouve dans des emplois qualificatifs : *les traditions (très) populaires* (que les gens apprécient) mais aussi dans des emplois relationnels : *une démocratie populaire* (du peuple). Certains emplois peuvent même être ambigus : *une chanson populaire* peut aussi bien être une chanson à succès qu'une chanson traditionnelle. Certains auteurs étendent le comportement relationnel à des adjectifs comme *rouge* ou *vert* (*l'armée rouge, la politique verte*). On ne peut donc pas, dans une perspective de classification sémantique, se concentrer sur l'adjectif lui-même, en ignorant la valeur sémantique du nom qu'il modifie, c'est-à-dire l'emploi dans lequel l'adjectif est engagé. Il est impossible de classer les adjectifs en classes étanches, comme on classe les unités lexicales en parties du discours. C'est pourquoi je propose de classer non pas les adjectifs en eux-mêmes, mais les emplois adjectivaux, un même adjectif pouvant apparaître dans différents types d'emplois. Goes montre ainsi que l'emploi relationnel est accessible à tout adjectif dénominal pourvu que le support nominal s'y prête. Le passage inverse est possible. Bartning et Noailly (1993) ont ainsi relevé pour *enfantin* toute une gamme d'emplois du relationnel pur au qualificatif pur:

- *Le langage enfantin* (des enfants),
- *Une émission enfantine* (pour enfants),
- *Une émotion enfantine* (qui a le caractère de l'enfance),
- *Une remarque enfantine* (caractéristique d'un enfant),
- *Un problème enfantin* (très facile).

Goes propose en conséquence de nuancer la différence entre relationnels et qualificatifs, et de parler d'adjectifs statistiquement relationnels ou statistiquement qualificatifs, en fonction du type d'emplois dans lequel l'adjectif s'engage plus volontiers. Les frontières entre les trois catégories d'emplois resteront cependant relativement floues. Romero note par exemple l'existence d'un continuum entre des emplois qualificatifs intenses comme *bouillant* dans *eau bouillante* et des emplois intensifs purs comme *énorme* dans *énorme envie*. *Echec cuisant, banalité consternante, amour passionnel* sont des exemples où l'adjectif est à la fois intensif

(une *banalité consternante* est une grande banalité) et qualificatif (consternant = qui est propre à consterner). L'intensité est plus ou moins présente dans les emplois qualificatifs. *Une étrange idée* ne comporte aucune intensification mais *une idée surprenante* est une idée qui surprend (qualification) parce qu'elle est très étrange (intensification). Quant à *l'eau glacée,* elle est somme toute *très froide* (intensification + qualification). Les emplois intensifs peuvent aussi contenir une part de qualification : *une rare hospitalité* intensifie *hospitalité* tout en la qualifiant (pas commune). Nous proposons donc, en accord avec Romero (2004), d'employer les termes d'emploi qualificatif, d'emploi relationnel et d'emploi intensif. Ce sont ces emplois que nous nous attacherons à caractériser au cours de notre exploration du lexique adjectival.

3. Le petit-monde des adjectifs

Nous l'avons dit, les graphes lexicaux, et plus généralement les grands graphes modélisant une activité humaine (réseaux sociaux, réseaux électriques, Internet…) dits « graphes de terrain » ou « réseaux réels », ont une structure commune très particulière dont les propriétés échappent aux modélisations classiques en théorie des graphes. Les graphes traditionnellement étudiés sont en effet 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 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 quelconques est faible en moyenne. Ce phénomène est surprenant mais non caractéristique d'une organisation. Erdös et Rényi (1959) 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 fortes chances 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 les graphes de terrain, c'est qu'ils sont peu denses, et possèdent à la fois une distance moyenne courte, comme les graphes aléatoires, et un fort taux de clustering, comme les graphes réguliers. 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 $C >> C_{aléatoire} \approx k/n$ alors que L est du même ordre de grandeur que $L_{aléatoire} \approx \ln(n)/\ln(k)$.

Les travaux de Watts et Strogatz ayant attiré l'attention sur les graphes de terrain, on a ensuite cherché à mieux les caractériser encore. Barabási et al. (1999) ont ainsi montré qu'ils font partie d'une autre classe très intéressante de graphes, les graphes **à invariance d'échelle**. Cela signifie que la répartition des degrés[2] des sommets suit une loi de puissance : la probabilité P(k) qu'un sommet du graphe considéré ait 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 à invariance d'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 à invariance d'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. Les premiers travaux menés sur les graphes de terrain 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. Gaume (2003) a ainsi été l'un des premiers en France 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

[2]Le degré d'un sommet est le nombre de sommets auquel il est relié, c'est à dire le nombre de voisins qu'il a dans le graphe.

connaissance de l'organisation du lexique. C'est dans le même esprit que nous travaillons.

Nous étudions le graphe du dictionnaire électronique des synonymes (désormais DES) du laboratoire CRISCO (www.unicaen.crisco.fr). 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. Le graphe correspondant est créé en reliant deux mots par un arc lorsqu'un des dictionnaires signale une relation synonymique entre eux. Le graphe correspondant possède 49 133 sommets et 198 549 arcs. Nous avons pu vérifier que ce graphe est un graphe de type petit-monde à invariance d'échelle. Avec ses 198 549 arcs pour 49133 sommets (donc de degré moyen k = 8.1), il est effectivement peu dense. On a ainsi :

L= 4.7306 (qui est bien du même ordre de grandeur du L d'un graphe aléatoire $L_{al} = \ln(49\,133)/\ln(8.1) \approx 5.1$)

C=0.35 (ce qui est très supérieur à ce qu'on aurait pour un graphe aléatoire, c'est-à-dire $C_{al} = 8.1/49\,133 \approx 1.6 \times 10^{-4}$

La distribution des degrés (Figure 2) ne suit pas vraiment une loi de puissance mais on a visiblement une structure '*hiérarchique'*. On a en fait seulement quatre sommets très connectés. Ce sont les mots *bon* (240 synonymes), *faire* (219 synonymes), *prendre* (210 synonymes) et *fort* (207 synonymes). La grande majorité des mots ont moins de 10 synonymes. Le nombre moyen de synonymes par mots est de l'ordre de 8, comme on l'a vu. La grande majorité des mots ont moins de 25 synonymes. De nombreux mots (14 985 au total) très spécifiques comme *abstentionnisme* ou *abscisse,* ou encore des noms propres comme *Jupiter* et *Cupidon* ne possèdent qu'un seul synonyme.

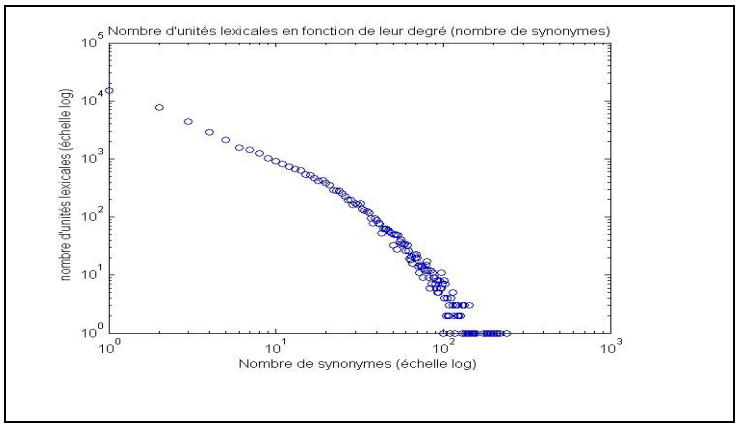

Figure 2. Répartion des degrés dans le DES

Ayant décidé de restreindre dans un premier temps notre étude au lexique adjectival, nous avons construit un graphe de synonymes adjectivaux, en croisant le DES avec les unités étiquetées comme adjectifs dans les sorties de l'analyseur Syntex (Bourigault et Fabre, 2000) sur un corpus constitué par tous les articles du journal *Le Monde* sur 10 ans. Ce graphe possède 3699 sommets et 22568 liens soit une moyenne de 6.10 synonymes par adjectif (nettement plus que pour le DES entier, où la moyenne est de 4.54), avec 2 adjectifs qui dépassent 150 synonymes (*beau* et *bon*), et 8 qui dépassent 100 (les 6 autres sont *dur, extraordinaire, fort, grand, mauvais* et *vif*). Sa composante connexe principale comporte 3614 nœuds et 22513 liens (les 85 synonymes écartés forment une kyrielle de petites composantes connexes de quelques éléments chacune). Elle est peu dense, la longueur maximale d'un chemin entre deux nœuds est de 14, et la longueur caractéristique L est de 4.04. Le coefficient de clustering C est de 0.28. Par comparaison avec un graphe aléatoire ayant le même nombre n de nœuds et le même nombre moyen de liens par nœuds k, pour lequel on a en moyenne : $L = \log(n)/(\log(k)) = 4.48$ et $C = k/n = 0.0017$, on voit que le graphe adjectival possède une structure de petit-monde.

4. Géométriser un graphe

4.1. *Utiliser des cliques*

Le but de l'exploration présentée ici est, nous l'avons dit, d'avoir accès à la topologie sous-jacente du graphe étudié. Nous nous appuyons pour cela sur sa structure de graphe petit-monde à invariance d'échelle. Ravasz et Barabási (2003) ont montré qu'un fort coefficient de clustering associé à une structure à invariance d'échelle déterminent une combinaison originale de modularité et d'organisation hierarchique. Un coefficient de clustering élevé traduit la présence de nombreux clusters, qui sont très interconnectés. Ces clusters s'associent entre eux pour former des groupes plus grands mais moins connectés, qui se combinent à nouveau pour former des clusters encore plus gros et encore moins connectés. Et ainsi de suite. L'invariance d'échelle fait que le nombre de sommets très connectés par rapport aux autres sommets est constante à chaque niveau de clusterisation. Cela fait qu'aucun sommet ne peut être vu comme dominant les autre. L'unité structurelle est donc le cluster : à la base on a des clusters vraiment petits et très connectés, qui deviennent de plus en plus gros et de moins en moins interconnectés. De plus, les groupes de sommets se recoupent à tous les niveaux. L'unité d'étude de la topologie du graphe doit donc être un petit groupe de sommets très fortement connectés les uns aux autres. C'est donc tout naturellement que, à la suite de Ploux et Victorri (98), nous utilisons la clique comme unité d'étude de notre graphe. Une clique est en effet un ensemble de sommets deux à deux connectés le plus grand possible. La Figure 3 présente un extrait du graphe adjectival. Ce graphe présente ainsi 3 cliques : < *bas ; brutal ; mauvais ; méchant* > (On ne peut pas ajouter *faible* à cette clique car il n'est

pas synonyme de *brutal*), < *bas ; faible ; mauvais* ; *méchant* > et < *fier ; intraitable ; méchant ; sauvage* >. La clique est un bon candidat de cluster de base dans la structure du graphe. Le taux de clusterisation élevé dans un graphe petit-monde assure la présence d'un grand nombre de cliques. L'utilisation des cliques comme unité minimale de groupements de sommets a par ailleurs fait ses preuves dans l'exploration de petits graphes de synonymie, constitués par un mot vedette et l'ensemble de ces synonymes. (Ploux et Victorri 98, Venant 04).

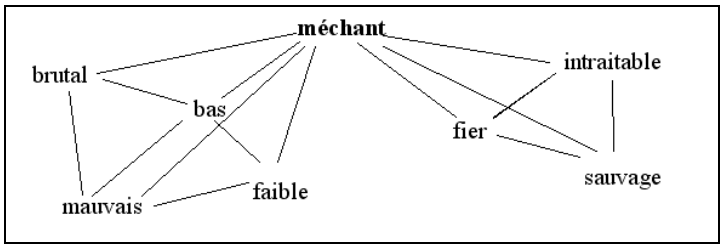

Figure 3. *Un extrait du graphe adjectival*

4.2. *Espaces sémantiques*

La structure hiérarchique que nous venons de décrire peut, la plupart du temps, être mise en relation avec un espace géométrique sous-jacent. Prenons l'exemple des réseaux sociaux, archétypes des graphes petit-monde à invariance d'échelle. Ces réseaux sont modélisés par des graphes dont les sommets sont des personnes. Les arêtes du graphe modélisent une relation sociale : être amis, connaître le prénom de l'autre personne, travailler ensemble… La structure d'un tel réseau est clairement reliée à un espace géographique sous-jacent, présentant une structure hiérarchique duale de celle du graphe. Les petits clusters de personnes très connectées correspondent à des petites zones géographiques où les gens se rencontrent régulièrement (bureaux, lieux d'habitation…). Ces lieux se regroupent ensuite en des zones géographiques plus étendues (société, quartiers, villages…) elles-mêmes très interconnectées (holdings, villes, communauté de communes...). L'hypothèse que nous faisons est que, à tout graphe de type petit-monde à invariance d'échelle, on peut associer un espace conceptuel sous-jacent, de nature souvent plus abstraite qu'une simple carte géographique, dont la topologie peut révéler celle du graphe, plus difficile d'accès. C'est cet espace que nous cherchons à construire par l'intermédiaire de l'utilisation des cliques. Dans l'exploration du graphe adjectival, il s'agira donc de construire l'espace sémantique associé au graphe. On peut considérer en première approximation qu'une clique correspond à un emploi adjectival et que ce sont donc les cliques qui constitueront les points de l'espace sémantique.

4.3 *Une métrique pour l'espace des cliques*

On peut définir l'espace sémantique adjectival comme l'espace euclidien engendré par les adjectifs. Chaque clique y est représentée par un point dont les coordonnées sont calculées en fonction des synonymes qu'elle contient : soient a_1, a_2, …, a_n les adjectifs, et c_1, c_2, …, c_p les cliques, l'adjectif a_i correspond au $i^{\text{ème}}$ vecteur de base de cet espace, et la clique c_k à un point dont les coordonnées x_{ki} valent 0 ou 1 suivant que l'adjectif correspondant appartient ou non à la clique : $x_{ki} = 1$ si $a_i \in c_k$ et $x_{ki} = 0$ si $a_i \notin c_k$.

La distance entre deux cliques c_k et c_l est alors donnée par la *métrique canonique* sur cet espace euclidien, définie de la façon suivante :

$$d^2(c_k, c_l) = \sum_{i=1}^{n} (x_{ki} - x_{li})^2$$

Ploux et Victorri (98), dans leur travaux sur la construction d'espaces sémantiques, montrent par l'analyse de quelques exemples que cette distance se révèle totalement inadéquate. Ils expliquent cela par le fait que cette distance donne le même « poids » à tous les synonymes, et qu'elle traite de la même manière toutes les cliques, quel que soit leur cardinal. « Or certains synonymes peuvent recouvrir une grande partie des emplois […], alors que d'autres sont plus « spécifiques », dans la mesure où ils ne s'appliquent qu'à un ensemble très restreint d'emplois. De plus, certaines cliques possèdent beaucoup plus d'éléments que d'autres. Ces différences doivent être prises en compte dans la définition de la distance, si l'on veut représenter correctement la proximité sémantique de deux cliques. » Ils proposent donc d'utiliser une métrique bien connue en analyse de données *la métrique du* χ^2 : deux cliques c_k et c_l étant données, la distance entre les deux est donnée par :

$$d^2(c_k, c_l) = \sum_{i=1}^{n} \frac{x}{x_{\bullet i}} \left(\frac{x_{ki}}{x_{k\bullet}} - \frac{x_{li}}{x_{l\bullet}} \right)^2$$

$$\text{Avec } x_{\bullet i} = \sum_{j=1}^{p} x_{ji} \text{ , } x_{k\bullet} = \sum_{i=1}^{n} x_{ki} \text{ , et } x = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ji} \text{ .}$$

Cette métrique possède l'avantage d'une part de pondérer chaque synonyme en fonction du nombre de cliques dans lequel il intervient (plus un synonyme apparaît dans des cliques différentes, moins il est spécifique et moins son rôle dans la discrimination des sens de l'unité est important), et d'autre part de diviser les coordonnées de chaque clique par son nombre d'éléments : le point représentant la clique est d'autant plus proche de l'origine que la clique correspondante comporte plus de synonymes. La distance du χ^2 confère donc à l'ensemble des cliques une

structure géométrique qui semble respecter la notion intuitive de proximité entre emplois.

5. A la découverte de l'espace sémantique adjectival

Nous avons calculé l'ensemble des cliques du graphe adjectival. Il en possède 11 900. La majorité des cliques ont entre 2 et 5 éléments, avec une moyenne de 4 éléments par clique. Pour comprendre la structure de l'espace adjectival nous n'avons pas immédiatement cherché à le visualiser dans sa totalité. Le grand nombre de cliques qui le constituent d'une part, et le grand nombre de dimensions qui l'engendrent d'autre part, rendent une telle visualisation peu manipulable. Pour avoir une idée de sa structure nous avons d'abord cherché à en explorer des morceaux, à nous promener à l'intérieur du nuage de points formé par les cliques. La construction de l'espace sémantique est telle que les cliques sont toutes à l'intérieur de l'hypercube unité[3]. Un premier moyen de comprendre comment s'organisent les cliques est de regarder leur répartition au sein de cet hypercube, et en particulier comment elles se situent par rapport à l'origine de l'espace. La Figure 4 montre l'histogramme des normes des vecteurs formés par les cliques (c'est-à-dire les distances à l'origine de chacune des cliques). Nous appellerons désormais *excentricité* d'une clique sa distance à l'origine.

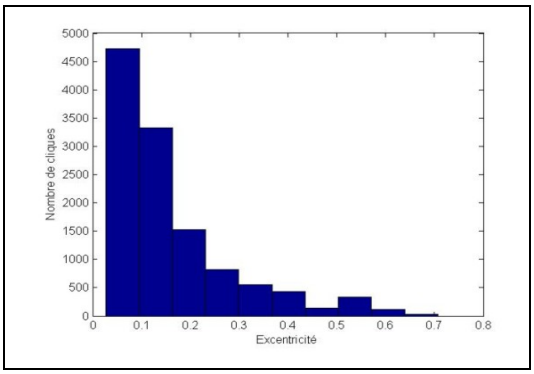

Figure 4. *Histogramme des excentricités adjectivales*

On voit que la majorité des cliques se situent à l'intérieur de la boule de rayon 0.2. On a aussi une forte densité de cliques dans la boule de rayon 0.1. L'espace devient de moins en moins dense en cliques au fur et à mesure qu'on s'éloigne de l'origine. On trouve quelques cliques très excentrées. La boule centrale est sans doute constituée des sens neutres ou généraux, puisque la distance du χ^2 est ainsi

[3] Un hypercube est la généralisation à plus de 3 dimensions de la notion de cube. En dimension n, c'est l'ensemble $\{ (x_1, x_2,\ldots, x_n) \in \Re^n , \max \{|x_1|, |x_2|,\ldots, |x_n|\} = 1\}$

faite que, pour être proche du centre, une clique doit contenir beaucoup de synonymes qui appartiennent à beaucoup de cliques. Les cliques centrales devraient ainsi correspondre au sens les plus désémantisés, et les cliques excentrées être des cliques plus courtes correspondant à des sens plus spécifiques.

5.1. *Exploration de la boule centrale*

Le logiciel VisusynGlobal utilisé pour la visualisation est une extension de Visusyn, le logiciel développé par Ploux et Victorri (98). Il nous permet, grâce à une analyse factorielle des composantes, de visualiser un ensemble quelconque de cliques extrait du nuage de cliques global. La grande densité de cliques aux alentours de l'origine nous a fortement incité à visiter d'abord cette région. Nous allons ainsi explorer la boule centrale de rayon 0.1. Elle contient 5036 cliques. Avec un si grand nombre de cliques une visualisation n'est pas forcément aisée à comprendre, aussi allons-nous procéder pas à pas, observant d'abord les cliques les plus proches du centre puis nous éloignant peu à peu jusqu'à visualiser la boule entière. Nous obtiendrons ainsi l'anatomie de notre boule un peu comme on dissèque un oignon : une fois la structure du bulbe comprise, il est plus facile de voir ce que chaque nouvelle couche apporte à la précédente.

Intéressons nous d'abord aux cliques les plus proches du centre, à savoir celles dont l'excentricité est inférieure à 0.04. Il y en a 245. Ce sont d'une part des cliques correspondant à des sens intensifs, d'autre part des cliques correspondant à des sens primaires. Deux nuances sont représentées dans les sens primaires. L'une a une valeur positive. Elle décline les sens de *beau* les plus généraux. L'autre a une valeur négative et concerne ce qui est *pénible* ou *mauvais*, en un mot *difficile*. Les sens intensifs se déploient sur deux pôles, l'un s'applique à des objets ou des événements, de façon objective, quant à leur taille ou leur ampleur (*colossal ; énorme ; extraordinaire ; fabuleux ; fantastique ; formidable ; gigantesque ; phénoménal ; prodigieux*). L'autre porte un jugement et qualifie plutôt l'effet produit sur le locuteur par l'évènement, l'objet ou la personne (*bizarre ; étonnant ; extraordinaire ; fantastique ; fantastique ; incroyable ; invraisemblable ; sensationnel ; surprenant*).

Lorsqu'on agrandit la boule et qu'on s'intéresse à l'ensemble des cliques dont l'excentricité est inférieure à 0.05 (cet ensemble contient le précédent), on voit apparaître un pôle intensif négatif avec des cliques comme *abominable ; affreux ; atroce ; épouvantable ; horrible ; mauvais.* Simultanément les sens primaires s'enrichissent. Du côté positif on voit apparaître des cliques comme *agréable ; aimable ; beau ; charmant ; enchanteur ; joli ; plaisant ; séduisant*. On commence un peu à s'éloigner des sens primaires pour entrer dans des nuances sémantiquement plus riches. Lorsqu'on agrandit le rayon de la boule centrale à 0.07, on voit apparaître un pôle d'intensifs s'appliquant à des caractères ou des comportements animés (*ardent ; bouillonnant ; enthousiaste ; exalté ; fanatique ; frénétique ; furieux*), un pôle d'intensifs négatifs généraux (*déplorable ; détestable ;*

lamentable ; méchant ; minable) ainsi que des branches organisant divers sens primaires (*austère ; bourru ; dur ; raide ; rude ; sec ; sévère*). Les sens déjà présents s'enrichissent aussi de nouvelles nuances.

Lorsqu'on visualise la boule de rayon 0.1 toute entière, on retrouve bien sûr les pôles décrits précédemment et on voit apparaître de nouvelles branches. La Figure 5 présente quelques unes de ces branches. Ces branches correspondent encore à des sens primaires comme *fade*, *dur* ou *juste*. Les sens des cliques de bout de branche correspondent cependant à des intensifications du sens général de la branche (avec des adjectifs comme *insignifiant*, *impitoyable* ou *certain).*

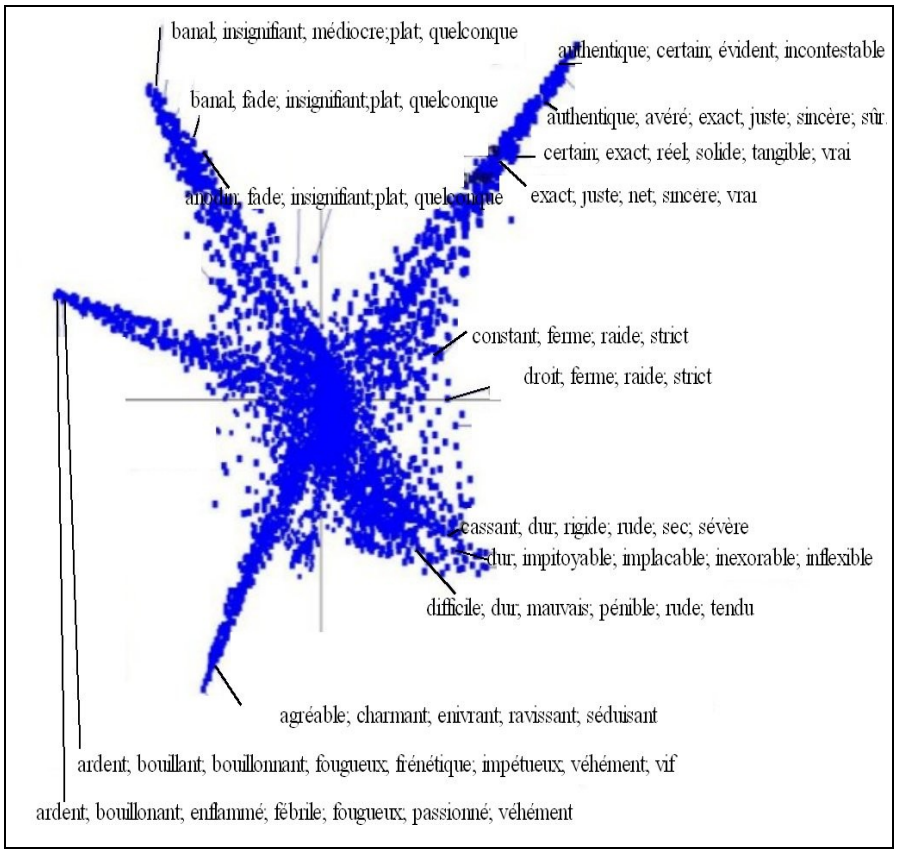

Figure 5 *Boule centrale de rayon 0.1.*

En résumé, on voit que les axes organisant la boule centrale sont de trois types : les premiers rendent compte d'une intensification, les seconds permettent d'organiser entre eux les sens primaires et les troisièmes opposent des valeurs positives à des

valeurs négatives. Les cliques présentes à l'intérieur de cette boule correspondent à des sens primaires (issus de la perception immédiate) ou intensifs. Ces sens sont très généraux voire désémantisés au centre de la boule. On s'éloigne de ce noyau dans toutes les directions en suivant des branches sémantiquement homogènes. Plus on s'éloigne du centre plus la coloration sémantique est grande.

On va maintenant s'intéresser à ces cliques, plus riches sémantiquement, et aux relations de proximité qu'elles entretiennent les unes avec les autres.

5.2. *Plus loin du centre*

Plus on s'éloigne du centre de l'espace, plus les cliques sont courtes et spécifiques. La longueur moyenne des cliques dont l'excentricité est supérieure à 0.1 est 3. Au-delà de 0.3, la taille maximale des cliques est 3 (comme *attendu ; désiré ; espéré* ou *affectif ; émotif ; émotionnel*), au-delà de 0.6 on n'a plus que des cliques à deux éléments (comme *russe ; soviétique*, *multilingue ; polyglotte* ou *femelle ; féminin*). Dans cette partie de l'espace, même les cliques les plus longues ont des sens très spécifiques. Les cliques ayant 8 sommets présentes dans cette partie de l'espace sont ainsi des cliques correspondant à des sens précis :

- bénéficiaire ; fructueux ; juteux ; lucratif ; payant ; profitable ; rémunérateur ; rentable

- éternel ; immortel ; immuable ; impérissable ; imprescriptible ; inaltérable ; indestructible ; sempiternel.

On veut en savoir un peu plus sur la façon dont est structurée cette partie de l'espace sémantique. Il ne suffit plus ici de visualiser l'ensemble des cliques. Nous avons vu en effet que les cliques s'organisent en branches sémantiquement homogènes. La structure de la boule centrale est assez simple et facile à repérer en faisant varier les axes de visualisation. Plus on s'éloigne du centre, plus les cliques correspondent à des sens précis, plus ces sens sont nombreux. L'omniprésence de la polysémie multiplie le nombre de branches. Ces branches se déploient dans toutes les directions, la projection en deux dimensions écrase cette structure et la rend parfois insoupçonnable. Il nous faut trouver un autre moyen de mettre en évidence les branches. Les branches rassemblent des cliques correspondant à des sens très proches mais pas forcément deux à deux (établissant une sorte de « ressemblance de famille » entre cliques). Pour les mettre en évidence, nous allons utiliser un outil géométrique, la boule, à partir duquel nous définissons et construisons des *branches* de cliques. :

Les boules : l'idée est de rassembler les cliques qui sont proches les unes des autres dans l'espace sémantique. Soit c une clique du graphe des adjectifs. On inclut dans la boule B de centre c toute clique c2 telle que d(c,c2)≤ r où

$$r^2 = \|c\|^2 + \min_{i=1}^{p}(\|c_i\|^2) .$$

On a choisi ce rayon de façon à exclure de la boule de centre *c* les cliques n'ayant aucun adjectif commun avec *c*. Rappelons qu'on a en effet :

$$d^2(c,c2) = \|c\|^2 + \|c2\|^2 - 2ps(c,c2).$$

Si l'intersection entre *c* et *c2* est vide, leur produit scalaire est nul et on a $d^2(c,c2) = \|c\|^2 + \|c2\|^2 \geq r^2$.

Les branches que nous cherchons à mettre en évidence sont des rassemblements de boules.

Les branches : On se fixe un seuil Se d'excentricité (distance à zéro). On forme C_{Se} l'ensemble des cliques d'excentricité supérieure ou égale à Se. Toute clique ***c*** de C_{Se} est potentiellement génératrice d'une branche BR_c. Cette branche contient B_c, la boule de centre ***c***. On va chercher à agrandir la branche BR_c en y incluant d'autres boules. Pour cela, on parcourt les éléments de C_{Se}. Pour chaque clique c_i de C_{Se}, on forme sa boule B_{ci}. Pour que B_{ci} soit incluse dans BR_c, il faut que le nombre de cliques qu'elle a en commun avec B_c, la boule de centre ***c***, soit au moins égal au cinquième du cardinal de la plus petite des deux boules. Une boule ne peut entrer que dans une seule branche. La formation de la branche s'arrête dès que l'intersection entre B_{ci} et B_c est vide. On crée alors une deuxième branche à partir de c_i .Le processus se poursuit jusqu'à ce que chaque boule ayant pour centre une clique de C_{Se} soit entrée dans une branche (et une seule). Le tableau 1 montre le nombre de branches obtenues en fonction du seuil d'excentricité choisi pour les cliques génératrices des branches.

Pour les valeurs les plus basses de Se les branches obtenues ne sont pas très significatives. C'est qu'on se trouve dans une zone très dense en cliques. Des sens très différents peuvent se trouver assez proches les uns des autres. Les points de départ des branches doivent être choisis dans des zones où les branches s'entrecoupent moins, des zones moins denses en cliques, plus excentrées. A partir de 0.3, on voit apparaître des branches plus cohérentes d'un point de vue sémantique. Certaines sont sémantiquement très précises. Elles peuvent s'appliquer plutôt à des êtres animés [4]:

< *abattu, bouleversé, brisé, chagrin, dépressif, déprimant, déprimé, désordonné, détruit, effondré, faible, fatigué, inerte, languissant, las, malade, marqué, mélancolique, miné, morne, morose, mou, piqué, ravagé, rompu, ruiné, saccagé, sombre, souffrant, tombé, tourmenté, travaillé, triste* >

Ou plutôt à des objets physiques :

[4]Bien que les branches soient des ensembles de cliques, nous n'indiquons ici que les adjectifs qu'elles contiennent.

< *accidenté, assimilé, bouclé, cassé, courbe, déchiqueté, découpé, gonflé, indirect, irrégulier, marqué, ondulé, plié, plissé, sinueux, soufflé, tordu, tortu, tortueux, tourmenté, tournant, varié* >

D'autres sont plus générales. La plus grosse branche rassemble 113 adjectifs dont le seul point commun est leur valeur négative (dans l'emploi correspondant) (*aberrant, affreux, destructeur, mortel...* mais aussi *lourd, léger, faible*).

Notons que l'on a une branche rassemblant des adjectifs de couleur :

< *blême, bleu, bleuté, céleste, cru, écarlate, interdit, mauve, noir, nouveau, pâle, pneumatique, pourpre, rouge, vert, violet* >

et des branches spécialisées dans les adjectifs relationnels :

< *administratif, bureaucratique, étatique, formaliste, gouvernemental, ministériel, officiel, public, réglementaire, tâtillon* >

ou encore

< *continental, français, hexagonal, Métropolitain, terrien, tricolore* >

Ou même

< *astral, céleste, cosmique, interplanétaire, interstellaire, lunaire, solaire, universel* >

Se	Nombre de branches	Nombre max d'adjectifs dans une branche
0.1	383	2896
0.2	599	660
0.3	648	113
0.4	554	55
0.5	405	55
0.6	78	7
0.7	26	2

Tableau 1: Tableau 1. Branches en fonction du seuil d'excentricité

Lorsqu'on augmente le seuil d'excentricité, les branches se spécialisent de plus en plus. Les plus grosses d'entre-elles peuvent cependant rester très générales. Ainsi la branche de taille maximale obtenue pour les seuils 0.4 et 0.5 est une classe rassemblant 55 intensifs positifs (*abracadabrant, admirable*…). Pour un seuil

d'excentricité supérieur à 0.5, la construction des branches est initiée dans des zones très peu denses en cliques ce qui explique que les branches soient très courtes et restent confinées dans les limbes de l'espace sémantique. La plus grosse branche obtenue pour Se = 0.6 contient 7 adjectifs :

< *aborigène, autochtone, indigène, justiciable, natif, naturel, ressortissant* >

Les branches obtenues rendent comptent de la polysémie à la manière dont les cliques le faisaient. Une clique, et a fortiori un adjectif, peut appartenir à plusieurs branches. On trouve l'adjectif *sec* dans 43 branches différentes, pour Se=0.4. Certaines correspondent à des sens primaires :

< *beau, bon, court, droit, facile, faible, gros, modeste, pauvre, petit, sec, sévère, simple, succinct, unitaire, vrai* >

D'autres ont des sens plus psychologiques :

<*avare, bourgeois, égoïste, entier, exclusif, indifférent, ingrat, insensible, intéressé, narcissique, personnel, sec* >

Remarquons sur le cas de l'adjectif *sec*, que l'exploration globale du lexique peut venir compléter utilement des informations locales. Lors d'un travail précédent (Venant, 2004), nous nous sommes intéressée au graphe formé uniquement par *sec* et ses synonymes. Dans ce graphe, le sens de *sec* dans *un atout sec* n'est représenté que par une seule clique, *sec ; seul ; simple*, ce qui posait des problèmes à notre système de désambiguïsation automatique. En travaillant sur le lexique adjectival global, pour un seuil Se=0.5, on trouve *sec* dans la branche :

< *abandonné, désert, distinct, indépendant, individuel, isolé, pur, retiré, sauvage, sec, seul, simple, singulier, solitaire, un, unique, veuf* >

Cette branche correspond au sens pris par *sec* dans *un atout sec*. Ce sens devient plus lisible au niveau global par le rapprochement de *sec* avec des adjectifs comme *isolé, solitaire* ou *singulier*.

6. Conclusion

La méthode d'exploration de graphes présentée ici donne de très bons résultats. Les visualisations mettent en évidence la structure sémantique du graphe adjectival. Nous savons désormais qu'il ressemble à une galaxie avec un noyau central très dense et des vides intersidéraux énormes. Ce sont les trois types sémantiques d'adjectifs relevés au paragraphe 2 qui organisent cette galaxie. Le noyau central contient uniquement des sens primaires et intensifs. Il est intéressant de noter que les sens les plus primaires, ceux qui se trouvent au cœur du noyau central, sont *beau*, *grand* et *mauvais,* ceux-là précisément que Wierzbicka (1996) considère comme des universaux sémantiques. De ce noyau central sont issues des branches sémantiques plus ou moins longues, très denses près du centre de l'espace et qui s'effilochent

ensuite dans toutes les directions. Ces branches s'entremêlent et se recoupent par le jeu de la polysémie. Elles ont cependant une grande homogénéité sémantique. Elles ont une coloration primaire, intensive, relationnelle ou qualificative. Les emplois relationnels se distinguent d'ailleurs clairement des emplois primaires et intensifs. Ils ont un sémantisme plus riche et sont donc plus périphériques. Notons quand même que si certaines cliques intensives sont très générales (celles contenant *colossal* ou *énorme* par exemple), d'autres sont beaucoup plus spécifiques (celles contenant *émérite* ou *pompeux* par exemple) et se regroupent dans des branches plus éloignées du noyau central. Les branches semblent aussi regrouper des sens de même extension. Certaines d'entre elles s'appliquent à des évènements ou des objets, d'autres à des êtres animés, d'autres enfin rassemblent des sens plus psychologiques. Le noyau rassemble les sens les plus désémantisés et généraux. Plus un sens est périphérique, plus son sémantisme est plein. Les visualisations obtenues recoupent les analyses linguistiques, tout en les éclairant d'un jour nouveau, et livrent de nouvelles informations sur les proximités sémantiques entre emplois adjectivaux. Des phénomènes sémantiques aussi particuliers que la désémantisation, ou le rôle central des sens primaires, peuvent être pris en compte.

Les expérimentations présentées ici apportent donc un éclairage théorique sur la catégorie des adjectifs. Elles nous ont permis, d'une part, de caractériser les grandes classes adjectivales traditionnellement distinguées: intensifs (Romero 2004), relationnels (Bally 1965, Bosredon 1988), qualificatifs et primaires (Goes 1999), et, d'autre part, de montrer que, d'un point de vue théorique, il ne fallait pas chercher à classer les adjectifs eux-mêmes, mais leurs emplois, un même adjectif pouvant appartenir à différentes classes suivant ses emplois. Cela est vrai y compris pour les adjectifs dits primaires. Même les adjectifs les plus basiques comme *beau* ou *grand* possèdent des sens très précis. La clique *beau ; cultivé ; intéressant* , par exemple, s'applique à des esprits, et la clique *beau ; correct ; élégant ; poli* s'applique à des individus. A l'inverse, les études de cas laissent penser que tout adjectif possède un ensemble de sens primaires représentés dans nos espaces sémantiques par ses cliques les plus centrales. Ces sens, qu'on peut donc qualifier d'emplois primaires, sont ceux pour lesquels les contraintes sémantiques sont réduites au minimum. On voit donc pourquoi la catégorie des adjectifs primaires est si difficile à délimiter : aussi loin qu'on aille dans la liste des adjectifs, qu'ils soient dérivés ou non, monosyllabiques ou pas, on trouvera toujours un sens de cet adjectif donnant à penser qu'il s'agit d'un adjectif primaire. De même on voit que tout adjectif s'éloigne, par ses emplois les plus périphériques, du prototype possiblement constitué par les adjectifs primaires. Goes (1999), dans sa recherche d'un prototype adjectival, s'est en effet d'abord intéressé à la catégorie des adjectifs primaires. On comprend son insatisfaction et la nécessité qu'il a eu ensuite de recourir à un prototype abstrait, ensemble de caractéristiques saillantes apparaissant le plus dans le plus grand nombre d'adjectifs. Notre exploration du lexique adjectival ne fait que commencer mais laisse apercevoir l'allure sémantique d'un tel prototype : des sens centraux très désémantisés, privilégiant éventuellement l'antéposition, puis une organisation sémantique en branches définie par des axes de type '*positif-négatif*', '*plus-ou-*

moins-intensif' ou encore '*plus-ou-moins-relationnel'*. Notons que l'organisation selon un axe '*positif-négatif*' avait été remarquée par Borodina (1963) sur certains adjectifs primaires de dimension (*grand, petit*), d'appréciation (*bon*, *mauvais*) ou encore de disposition personnelles (*brave, lâche*). Il serait d'ailleurs intéressant d'explorer plus avant la boule centrale pour voir si on y trouve trace de l'organisation en huit classes sémantiques proposée par Borodina : adjectifs de dimension, (*grand, petit, haut, bas*), adjectifs de temps (*bref, vieux, jeune*), adjectifs d'appréciation (*bon, mauvais, joli, cher),* adjectifs de couleur, adjectifs de propriété physique (*chaud, froid, beau, laid),* adjectifs modaux (*vrai, faux*), adjectifs de disposition personnelle *(fort, faible, brave, lâche)* et adjectifs de vitesse (*rapide, lent, leste*). Notons enfin que le fait de pouvoir décider automatiquement si un emploi adjectival est sémantiquement riche ou non, en fonction de la distance au centre de l'espace de la clique correspondante, peut être d'une grande utilité dans le traitement des changements de sens entre anté et postposition par un système de désambiguïsation. Il s'agit par exemple d'attribuer à *méchant* antéposé son sens intensif de « déficient, de mauvaise qualité, médiocre… » (*un méchant cheval, un méchant costume*) alors que postposé, il prend des sens plus riches sémantiquement : *un enfant méchant, un rire méchant* (cf. Venant, 2006). La caractérisation des emplois adjectivaux que nous proposons (relationnels, intensifs, primaires), utilisée dans une tâche de désambiguïsation, trouve une application directe dans la détection de l'analyse automatique des opinions, car elle peut permettre, par exemple, d'évaluer la subjectivité des adjectifs en contexte, et de rendre compte du continuum qui existe entre les emplois adjectivaux objectifs (*un homme célibataire*, *un costume féminin*) et les emplois adjectivaux subjectifs (*un bel homme, un costume étrange*). La caractérisation d'emplois adjectivaux peut s'avérer aussi particulièrement intéressante dans la problématique de la mise à jour de thesaurus ou de la veille scientifique « où l'apparition d'un adjectif relationnel semble caractériser une stabilisation d'un concept scientifique émergeant. » (Daille 99). Etendue à l'ensemble du lexique, cette exploration devrait permettre de mettre au jour des axes organisateurs du lexique, (concret/abstrait, positif/négatif...) qui seront très utiles dans les analyses axiologiques (catégorisation pour/contre de critiques de cinéma apparaissant dans un forum, par exemple).

Cette exploration ne révèle cependant qu'un certain aspect de la structure lexicale, puisque nous ne l'avons abordée que par le biais des relations de synonymie. Cette vue, partielle, est à compléter par l'étude d'autres relations lexicales qu'elles soient sémantiques (antonymie, hyperonymie...) ou non (dérivation, suites syntaxiques: adjectif - nom, verbe – adverbe..., rapports syntaxiques: verbe - nom(sujet), verbe – nom(objet), ...). Les travaux que nous avons menés en calcul du sens (Venant 2006) ont montré l'intérêt de coupler renseignements externes et internes au corpus, surtout dans le cadre de la recherche d'information. La désambiguïsation que nous avons mise en place fait intervenir ces deux types de connaissances, venant de la langue générale ou spécifique au corpus.

La prochaine phase consistera donc à tenter d'explorer d'autres graphes lexicaux. On peut espérer obtenir des résultats aussi intéressants, pour peu que ces graphes soit de type petit-monde à invariance d'échelle, ce qui est souvent le cas. Nos outils offrent aussi des perspectives quant à l'exploration du Web, en particulier du Web sémantique quand il existera vraiment. Cela suppose de construire le graphe induit par les liens contenus dans les pages Web. Une clique dans ce cas correspondra à un sujet commun, assez précis, dont parlent toutes les pages concernées. On peut, puisque le Web possède une structure petit-monde (Barabási *et al.* 2000), supposer qu'on obtiendra une structure analogue à celle du lexique adjectival: un noyau central regroupant les sujets les plus courants et des branches rayonnant depuis ce noyau et qui organisent les grands thèmes discutés sur le Web. Il pourrait alors être intéressant de suivre l'évolution d'une branche : comment elle naît, se développe et parfois meurt.

8. Bibliographie

Barabási A-L, A R., Jeong H., « Scale free characteristics of random networks: The topology of the World Wide Web ». *Physica A*, 281:69.77, 1999.

Barabási A.-L., Albert R., Jeong H. and Bianconi G., « Power-Law Distribution of the World Wide Web », *Science*, 287:2115a, 2000.

Barque L., « FrameNet et le traitement de la polysémie », *Journée d'étude de l'ATALA*, 13 mai 2006

Bartning I. et Noailly M., « du relationnel au qualificatif : flux et reflux », *L'information Grammaticale*, 58, L'adjectif (M. Noailly ed.), 1993.

Bouaud J, Habert B., Nazarenko A et Zweigenbaum P., « Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation à deux modélisations conceptuelles », *Ingénierie des Connaissances, évolutions récentes et nouveaux défis*, J. Charlet, M. Zacklad, G. Kassel and D. Bourigault, (Eds). Eyrolles, 2000.

Bourigault D., Fabre C., « Approche linguistique pour l'analyse syntaxique de corpus », *Cahiers de Grammaires,* n° 25, Université Toulouse - Le Mirail, 2000 pp. 131-151.

Borodina M.A., « L'adjectif et les rapports entre sémantique et grammaire en français moderne », dans *Le Français Moderne*, XXXI-3, p. 193-198., 1963

Daille B., « L'identification en corpus d'adjectifs relationnels : une piste pour l'extraction automatique de terminologie », *TAL, Volume 42 Lexiques sémantiques*, 2001.

Daille B., Identification des adjectifs relationnels en corpus, TALN'99, ATALA, p. 105-114, Cargèse, Corse,1999.

Bally C., *Linguistique générale et Linguistique française*, 4ème ed., Berne, Francke, 1965.

Bosredon A. (1988), Un adjectif de trop, l'adjectif de relation, *L'information grammaticale*, n°37, 1988

Erdös P. and Rényi,A., « On Random Graphs I », *Publ. Math. Debrecen*, vol. 6, 1959

Felbaum C., *WordNet An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

Gaume 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., Hathout N. et Muller P., « Word Sense Disambiguation using a dictionary for sense similarity measure »,. *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004),* volume II, pp. 1194-1200. Genève, Suisse.

Goes J., *L'adjectif entre nom et verbe,* Paris – Louvain –La - Neuve, Duclot, 1999.

Noailly M., *L'adjectif en français moderne*, Paris, Ophrys., 1999.

Ploux S., Victorri B., « Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes », *Traitement Automatique des Langues*, 39(1):161-182, 1998.

Pottier B., De l'adjectif, *Travaux de Linguistique et le Littérature.*, XXIII-1, 1985.

Ravasz E., Barabási A.L. « Hierarchical Organization in Complex Networks », *Phys. Rev. E*, 67, 026112, 2003.

Reiner E., *La place de l'adjectif épithète en français : théories traditionnelles et essai de solution*, Wien, Stuttgart, W. Braumuller, Band, 1968.

Romero C., « Les adjectifs intensifs », *François J. l'adjectif en français et à travers les langues*, Presses Universitaires de Caen, 2004.

Roget P. M., *Roget's International Thesaurus*, Collins, Hardcover, 1963.

Véronis J., « Cartographie lexicale pour la recherche d'information », *Actes de la conférence TALN,* 2003.

Venant F., « Polysémie et calcul du sens », *Le poids des mots, Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles (JADT),* 2004.

Venant F., Représentation et calcul dynamique : exploration du lexique adjectival du français, mémoire de doctorat de l'EHESS, 2006.

Watts D.J. and Strogatz S. H., « Collective dynamics of "small – world" networks », *Nature*, 393, 1998.

Wierzbicka A., *Semantics: primes and universals*. Oxford:Oxford University Press, 1996.