Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes

Sabine Ploux, Bernard Victorri

Traitement automatique des langues, 39, n°1, 1998 – Version préliminaire

Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes

Sabine Ploux, Bernard Victorri

Introduction

La description sémantique des unités lexicales est un enjeu important pour le traitement automatique des langues. Comme on le sait bien, ce n'est pas un problème simple. Cela est dû en grande partie à l'omniprésence de la polysémie, qui touche beaucoup d'unités de la langue, et en premier lieu, les mots les plus usuels : en général, plus une unité est utilisée couramment, plus elle présente d'acceptions différentes, et plus sa structure sémantique est complexe[1].

Les approches classiques (analyse sémique, décomposition en primitives, réseaux sémantiques hiérarchiques, etc.), outre qu'elles ne sont pas particulièrement adaptées à traiter les problèmes posés par la polysémie, sont forcément limitées dans leur ambition : elles réclament un travail d'analyse considérable pour chaque unité étudiée, et elles sont souvent restreintes à un domaine spécifique, ce qui rend difficilement envisageable le traitement de tout le lexique par ces méthodes.

Aussi se tourne-t-on de plus en plus vers des méthodes automatiques ou semi-automatiques, rendues possibles par le progrès technologique qui permet aujourd'hui de disposer facilement de nombreuses ressources textuelles : corpus de grande taille, dictionnaires électroniques, etc. L'idée générale est d'utiliser ces ressources pour décrire la sémantique des unités à partir d'analyses statistiques des relations paradigmatiques et syntagmatiques qu'elles entretiennent les unes avec les autres.

On peut distinguer deux types de travaux qui vont dans cette direction (pour une présentation générale récente, cf. Habert B. *et al*. 1997). Certains travaux opèrent avant tout sur l'axe syntagmatique grâce à l'analyse de corpus (cf. entre autres Hindle D. 1990, Grefenstette G. 1994) : chaque unité est caractérisée sémantiquement par l'ensemble des unités avec lesquelles elle entre en relation syntaxique. D'autres travaux (cf. entre autres, Veronis J. & Ide N. 1990, Warnesson I. 1992) se placent directement sur l'axe paradigmatique en utilisant les données de dictionnaires électroniques, et, en particulier, en se servant de la relation de synonymie qu'ils permettent de mettre en évidence.

L'étude que nous présentons ici s'inscrit dans cette dernière lignée. Il s'agit en effet d'utiliser des dictionnaires de synonymes pour décrire la structure sémantique d'unités lexicales. Comme on va le voir, la principale originalité de ce travail consiste en l'utilisation du cadre mathématique du continu, un espace sémantique étant associé à chaque unité lexicale. Nous allons donc d'abord préciser les bases linguistiques et le modèle mathématique sur lesquels se fondent notre travail, avant de présenter les outils informatiques que nous avons mis au point et les premiers résultats que nous avons obtenus.

---

1. Cette constatation vaut bien sûr autant pour les *unités grammaticales* (appartenant à des classes « fermées » : déterminants, prépositions, conjonctions, etc.) que pour les *unités lexicales* proprement dites (appartenant à des classes « ouvertes » : noms, verbes, adjectifs, etc.). Nous ne nous intéressons ici qu'aux unités lexicales, les unités grammaticales méritant à notre avis des traitements plus spécifiques.

1. Polysémie et synonymie

1.1. Définition de la synonymie

Commençons d'abord par bien préciser la notion de synonymie que nous utilisons ici. Il existe en fait deux relations différentes que l'on nomme synonymie, et qui n'ont pas les mêmes propriétés : la synonymie « pure » et la synonymie « partielle ». On peut définir la synonymie pure de la manière suivante :

Deux unités lexicales sont en relation de synonymie pure si toute occurrence de l'une peut être remplacée par une occurrence de l'autre dans tout environnement sans modifier notablement le sens de l'énoncé dans lequel elle se trouve.

Par *environnement*, on entend à la fois le *co-texte* (ensemble des unités linguistiques présentes dans l'énoncé et au delà, dans le texte) et le *contexte* (ensemble des conditions d'énonciation et de la situation extra-linguistique)[2].

Telle que nous venons de la définir, la relation de synonymie pure est une relation d'équivalence : elle est réflexive, symétrique et transitive[3]. Elle permet donc de construire des classes d'unités lexicales regroupant les synonymes. Mais cette définition est très restrictive. Il existe très peu (sinon pas du tout) de synonymes purs, en raison même de l'omniprésence de la polysémie : la plupart des unités n'ayant pas le même sens d'un énoncé à l'autre, il serait étonnant qu'elles restent paraphrasables par les mêmes autres unités dans tous les environnements.

Aussi utiliserons-nous ici la notion de synonymie partielle, que nous appellerons désormais *synonymie* tout court, que l'on peut définir de la façon suivante :

Deux unités lexicales sont en relation de synonymie si toute occurrence de l'une peut être remplacée par une occurrence de l'autre dans un certain nombre d'environnements sans modifier notablement le sens de l'énoncé dans lequel elle se trouve.

Cette relation est toujours réflexive et symétrique, mais elle n'est plus transitive : une unité peut être paraphrasable par une première unité dans certains environnements, et par une deuxième unité dans d'autres, sans que ces deux unités soient elles-mêmes synonymes : il suffit pour cela que les deux ensembles d'environnements en question soient disjoints. Pour ne prendre qu'un exemple, *défendre* et *interdire* sont synonymes parce qu'ils sont paraphrasables l'un par l'autre dans des énoncés tels que *défendre de fumer* et *interdire de fumer*. De même, *défendre* et *soutenir* sont synonymes parce que *défendre les droits de l'homme* et *soutenir les droits de l'homme* ont sensiblement le même sens. En revanche, *interdire* et *soutenir* ne sont pas synonymes car il n'existe pas d'environnement dans lesquels on puisse les permuter sans modifier considérablement le sens de l'énoncé dans lequel ils se trouvent.

Cette définition de la synonymie permet à son tour de définir « proprement » les notions de polysémie, de monosémie et d'homonymie :

Une unité lexicale est dite monosémique si tous ses synonymes sont synonymes entre eux[4].

Une unité lexicale est dite homonymique si l'ensemble de ses synonymes (autres qu'elle-même) est séparable en au moins deux sous-ensembles disjoints pour la relation de synonymie,

---

2. Cette définition dépend aussi bien sûr de ce que l'on entend précisément par « le sens d'un énoncé » et par « modification notable de ce sens ». Voir Victorri B. & Fuchs C. 1996 chap. 1 pour une présentation cohérente d'un cadre théorique général assez classique dans lequel peuvent s'inscrire ces définitions.

3. Rappelons qu'une relation R sur un ensemble E est dite *réflexive* si a R a pour tout $a \in$ E, *symétrique* si a R $b \Rightarrow b$ R a pour tout a et $b \in$ E, et *transitive* si (a R b et b R c) $\Rightarrow a$ R c pour tout a, b et $c \in$ E. Une relation d'équivalence sur un ensemble E permet de faire une partition de E en *classes d'équivalence*, en regroupant les éléments équivalents.

4. Attention : cela ne signifie pas que cet ensemble de synonymes est fermé pour la relation de synonymie. N'importe lequel d'entre eux peut avoir d'autres synonymes qui ne soient pas des synonymes de l'unité monosémique considérée.

c'est-à-dire tels que chacun des éléments de l'un des sous-ensembles n'est synonyme d'aucun des éléments des autres sous-ensembles.
Remarque : On préfère généralement considérer que l'on a alors affaire à *des* homonymes, c'est-à-dire à plusieurs unités lexicales (non synonymes entre elles) partageant une même forme, en associant à chacune de ces unités l'un des sous-ensembles disjoints de synonymes.

Enfin *une unité est dite polysémique si elle n'est ni monosémique ni homonymique, c'est-à-dire si elle admet des synonymes qui ne sont pas synonymes entre eux, mais qui sont toujours reliés par synonymie entre eux à l'aide d'une chaîne de synonymes de l'unité considérée (et différents de cette unité)*[5].
Pour reprendre notre exemple, *défendre* est polysémique selon cette définition. En effet, même si *soutenir* et *interdire* sont loin d'être synonymes, il existe des chaînes de synonymes de *défendre* qui les relient, comme la chaîne suivante :

soutenir ↔ *protéger* ↔ *garder* ↔ *tenir* ↔ *empêcher* ↔ *interdire*.

En fait, la polysémie ainsi définie apparaît comme le cas général, la monosémie et l'homonymie représentant des cas extrêmes opposés.
Ainsi la structure conférée à l'ensemble des unités lexicales par la relation de synonymie permet de caractériser un certain nombre de propriétés sémantiques de ces unités. Mais pour aller plus loin dans leur description sémantique, il faut faire des hypothèses supplémentaires sur la nature du phénomène de la polysémie : selon le cadre théorique dans lequel on se placera, l'utilisation que l'on fera de la relation de synonymie sera différente.

1.2. Les approches discrètes de la polysémie

Dans les approches discrètes de la polysémie, on considère que l'on peut associer à chaque unité polysémique un nombre fini de « sens », exclusifs les uns des autres. La relation de synonymie peut alors être caractérisée de la façon suivante : deux unités sont synonymes si elles partagent au moins un de leurs sens. Ainsi, si l'unité u_1 possède les n sens s_1^1, s_1^2, ..., s_1^n, et l'unité u_2 les m sens s_2^1, s_2^2, ..., s_2^m, u_1 et u_2 sont synonymes s'il existe au moins un couple (i,j) tel que $s_1^i = s_2^j$.
Il s'agit donc en fait d'une réduction de la polysémie à une « homonymie généralisée » : on peut en effet dire de manière équivalente que chaque unité lexicale se décompose en un certain nombre d'unités homonymes, et que la relation de synonymie (partielle) entre unités lexicales se réduit à une relation de synonymie pure au niveau plus fin où l'on a décomposé ces unités. Par exemple, *défendre* est décomposé en $défendre_1$, $défendre_2$, etc., *interdire* en $interdire_1$, $interdire_2$, etc., et la relation de synonymie entre *défendre* et *interdire* provient d'une relation de synonymie pure entre, disons, $défendre_2$ et $interdire_1$. Cette relation de synonymie pure est, comme on l'a vu, une relation d'équivalence. Les « sens » sont alors définis comme les classes d'équivalence issues de cette relation : ainsi la classe d'équivalence ($défendre_2$, $interdire_1$, $prohiber_1$, etc.) définit un sens différent de la classe d'équivalence ($défendre_1$, $soutenir_3$, $garantir_2$, etc.).
C'est cette approche qui a été choisie dans WordNet (cf. Miller G.A. *et al.* 1993) : les classes d'équivalence y sont appelées des *synsets* et les sens associés des *concepts*. De même, le travail comparable mené sur le français depuis longtemps par l'équipe de Mémodata (dont le produit de base s'appelle *Dicologique*[6]) est fondé, à des nuances près, sur une conception similaire de la polysémie.

---

5. Attention là encore : il est essentiel de se limiter dans ces définitions à l'ensemble des synonymes de l'unité étudiée (en excluant celle-ci, bien entendu). En effet, même dans les cas les plus indiscutables d'homonymie, on trouve souvent une chaîne extérieure à cet ensemble qui relie deux synonymes appartenant à des sous-ensembles disjoints. D'une manière générale, la clôture transitive de la relation de synonymie est sans intérêt, parce qu'elle conduit à des classes d'équivalence extrêmement vastes, constituées d'unités aux sens les plus divers.
6. On trouvera une description de *Dicologique* sur le site Web de Mémodata : http://www.memodata.com. Voir aussi Dutoit D. 1992.

Peut-on utiliser les dictionnaires existants de synonymes, qui, comme on le verra, s'appuient sur la relation de synonymie partielle entre unités lexicales, pour construire automatiquement ces classes d'équivalence ? Le travail d'Isabelle Warnesson (Warnesson I. 1985, 1992) s'inscrit dans cette problématique. Sans entrer dans les détails ici, disons qu'elle propose un traitement en deux étapes :
- une « désambiguïsation » des unités lexicales, obtenue par un algorithme d'agrégation incrémentale, qui permet de séparer grossièrement les sens des unités polysémiques, et d'attribuer la relation synonymique donnée par le dictionnaire aux sens appropriés ainsi dégagés.
- la construction d'une relation d'équivalence entre ces sens, obtenue en calculant (par une technique de programmation linéaire) la relation d'équivalence la plus proche (en un sens très précis) de la relation obtenue précédemment.
Les résultats obtenus par cette méthode sont très encourageants. Ils montrent tout l'intérêt de cette voie de recherche : il semble possible, dans le cadre d'une approche discrète de la polysémie, d'utiliser les dictionnaires de synonymes pour automatiser la description sémantique des unités lexicales.

1.3. Vers une approche continuiste de la polysémie

Les approches discrètes reposent sur une hypothèse extrêmement forte : il faut postuler l'existence de ce niveau de « sens » ou de « concepts » qui permet de catégoriser les emplois d'une unité lexicale polysémique. Il est clair que cette hypothèse a de grandes vertus pour les applications concrètes. En réduisant ainsi la complexité du phénomène de la polysémie, on se donne un cadre de travail efficace qui suffit probablement pour un grand nombre d'applications, où le problème n'est pas tant d'analyser finement les sens d'une unité lexicale, que d'éviter les erreurs grossières auxquelles la polysémie peut conduire. Ainsi le fait de distinguer pour *défendre* un sens « positif » (*soutenir*, *garder*, etc.) et un sens « négatif » (*interdire*, *empêcher*, etc.) suffit amplement pour les besoins de nombreuses applications. L'approche discrète, même grossière, peut donc rendre de très grands services.
Il faut cependant remarquer que cette hypothèse est très simplificatrice et qu'elle ne repose pas sur une base théorique solide. Dès que l'on étudie de plus près telle ou telle unité lexicale, on se rend compte que cette approche conduit à de nombreuses difficultés. Il y a une grande part d'arbitraire, qui semble irréductible, dans la détermination de ce niveau de « sens ». Combien de sens, par exemple, pour *défendre* ? Deux, comme nous l'avons suggéré, semble nettement insuffisant pour une description qui cherche à cerner un tant soit peu la sémantique de ce verbe : il n'y a pas qu'une « nuance » de sens entre, par exemple, *défendre un assassin*, *défendre une conception du monde* et *défendre un bastion* ou entre *se défendre de fumer* et *se défendre d'avoir fumé*. Alors combien ? cinq ? dix ? vingt ? Sur quels critères prendre une décision ? Les méthodes automatiques, comme celle de Warnesson, peuvent décider à notre place. Mais comment juger de l'adéquation des résultats si l'on n'a pas de critère théorique préétabli ? En fait, le problème provient de l'existence d'un grand nombre de cas intermédiaires, qui interdisent de tracer des frontières sémantiquement fondées : *défendre l'entrée d'un port de guerre*, c'est à la fois en *interdire* l'accès à des agresseurs éventuels, et en *garantir* l'accès aux forces alliées. L'importance respective de ces deux fonctions dépend plus de l'état des tensions internationales que de la langue… Si l'on opte pour un petit nombre de sens, ces cas intermédiaires se trouveront « à cheval » sur une frontière. Si au contraire on choisit une division plus fine, la plupart des emplois engloberont plusieurs sens à la fois.

Ce problème peut être résolu si l'on change de cadre théorique, en adoptant une approche continuiste de la polysémie. En effet, dans ce cadre (pour une présentation détaillée voir Victorri B. & Fuchs C. 1996), on associe à chaque unité polysémique un *espace sémantique*, de petite dimension, muni d'une structure mathématique précise, et l'on représente le sens de l'unité dans chacun de ces emplois par une *région* de cet espace sémantique. Ce modèle présente un certain

nombre d'avantages. D'abord, le fait de plonger les différents sens d'une unité dans un même espace muni de propriétés adéquates permet de rendre compte des relations intuitives de voisinage entre ces sens, qui sont à l'œuvre dans les jugements spontanés des locuteurs (qui s'expriment par des réflexions telles que *ces deux sens sont très proches*, *là on s'éloigne nettement du sens initial*, etc.). Ensuite, le fait de représenter les sens par des régions permet de modéliser différents cas de figure dans les emplois de l'unité : un sens très « précis » sera représenté par une petite région de l'espace sémantique, un sens intermédiaire plus indéterminé par une région englobant plusieurs sens précis, une ambiguïté-alternative par une région composée de deux (ou plusieurs) sous-régions non connexes, etc. Enfin, et ce n'est pas le moindre des avantages, ce modèle explique les difficultés de l'approche discrète : réduire la polysémie à un nombre fini d'homonymes, c'est faire une partition de l'espace en autant de zones. Si l'on choisit une partition en un petit nombre de « grandes » zones, les régions représentant les sens intermédiaires seront en partie sur plusieurs zones. Si l'on choisit une partition très fine, ces régions engloberont plusieurs « petites » zones. Avec ce modèle, on peut donc rendre compte de toute la complexité du comportement sémantique des unités lexicales, tout en travaillant dans un espace de dimension raisonnable (c'est-à-dire décrit par un petit nombre de paramètres) : ce sont les relations topologiques entre régions qui sont complexes, et non l'espace dans lequel ces régions se déploient.

Pour qui s'intéresse donc à des descriptions fines des unité lexicales, que ce soit d'un point de vue théorique de modélisation du comportement sémantique des unités de la langue, ou d'un point de vue applicatif, pour des réalisations qui réclament de manière impérative une représentation précise du sens lexical, l'approche continue de la polysémie constitue une voie intéressante, qui mérite d'être exploitée plus largement qu'elle ne l'a été jusqu'à présent.

La question qui se pose alors, c'est de savoir si la relation de synonymie peut être utilisée dans le cadre de l'approche continue pour construire automatiquement l'espace sémantique associé à une unité lexicale. C'est à la réponse à cette question qu'est consacré le reste de cet article.

2. Graphe de synonymes et représentation du sens

2.1. Construction d'un graphe de synonymes

Les ressources dont nous disposions au départ de ce travail[7] étaient constituées de listes de synonymes extraites de sept dictionnaires français : le *Bailly*, le *Benac*, le *du Chazeaud*, le *Guizot*, le *Lafaye*, le *Larousse*, et le *Robert*. Ces dictionnaires sont très disparates, tant par leur couverture que par leurs objectifs. Nous avons considéré qu'au delà des intentions des auteurs de ces dictionnaires, la présence d'un terme dans l'une quelconque de ces listes suffisait à indiquer que, dans certains environnements, le mot-vedette et le terme cité étaient permutables sans modification notable du sens des énoncés en question. Conformément à notre définition de la synonymie, cela suffisait donc à valider une relation de synonymie (partielle) entre eux. Nous avons donc procédé à la fusion de ces listes (en regroupant en une seule liste les listes issues des différents dictionnaires pour une même entrée), puis à leur symétrisation (chaque fois qu'une entrée u_1 comporte une unité u_2 dans sa liste, on rajoute u_1 à la liste associée à l'entrée u_2 si elle ne s'y trouve pas déjà)[8].

Nous avons ainsi obtenu un dictionnaire[9] d'environ 40 000 entrées, nettement plus riche que chacun des dictionnaires sources. Il faut noter cependant qu'il y a eu aussi perte d'informations : dans la plupart de ces dictionnaires, les articles sont structurés en sections et sous-sections regroupant les

7. Ce travail est le fruit d'une collaboration entre l'INaLF et l'ELSAP. C'est l'INaLF qui nous a fourni les données, et qui est à l'origine de l'idée d'utiliser les cliques (voir plus bas) pour différencier les sens des unités lexicales. Nous tenons à remercier Robert Martin pour le soutien attentif et avisé qu'il nous a manifesté tout au long de cette recherche.

8. On trouvera des détails et des exemples de ces opérations dans Ploux S. *à paraître*.

9. Un grand nombre d'erreurs, typographiques notamment, ont aussi été corrigées au cours de ce travail par Jean-Yves Lacroix à l'ELSAP (plus d'un millier de corrections ont déjà été effectuées).

sens voisins. Mais d'une part ces regroupements sont très variables d'un dictionnaire à l'autre, ce qui rend difficile, voire impossible, leur mise en cohérence par des moyens automatiques dans un dictionnaire fusionné. Et d'autre part, ces regroupements obéissent à des principes de hiérarchisation des sens difficilement conciliables avec notre objectif : nous voulions justement voir dans quelle mesure l'utilisation de la seule relation de synonymie permettait de structurer les sens d'une unité polysémique, sans préjuger a priori de la possibilité d'obtenir une structure hiérarchique. C'est donc sciemment que nous n'avons pas pris en compte ces informations supplémentaires dans les traitements.
Ce nouveau dictionnaire[10] ne fait donc que donner de manière explicite le graphe de la relation de synonymie (selon la définition que nous avons donnée de cette relation) pour les unités lexicales de la langue française, du moins telle qu'on peut la déduire des données des sept dictionnaires sources que nous avons utilisés. C'est uniquement sur ce graphe que repose l'ensemble des traitements que nous allons maintenant présenter.

2.2. La notion de clique

C'est la notion de *clique* qui est au centre de notre travail sur le graphe des synonymes. Une clique (sous-graphe complet maximal[11]) est un ensemble le plus grand possible de sommets du graphe tous reliés deux à deux, « le plus grand possible » voulant dire qu'il n'existe pas de sommet n'appartenant pas à l'ensemble qui soit relié à tous les sommets de l'ensemble. En d'autres termes, pour notre application, une clique est un ensemble d'unités lexicales qui ont la propriété d'être toutes synonymes les unes des autres, aucune autre unité ne pouvant être rajoutée à l'ensemble sans que l'on perde cette propriété.
Prenons, à titre d'illustration, le graphe de la figure 1, dont les sommets u_1, u_2, …, u_{10} représentent des unités lexicales, et dont les arêtes indiquent la relation de synonymie.

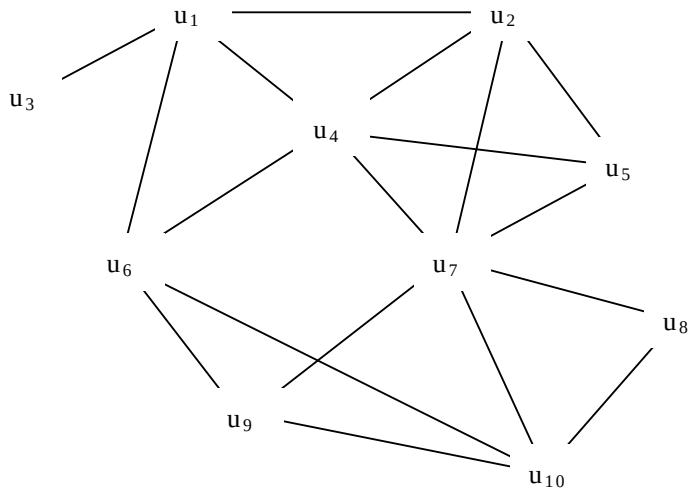

Figure 1 : Exemple de graphe

Les ensembles $\{u_1, u_2, u_4\}$, $\{u_2, u_4, u_5, u_7\}$, $\{u_1, u_3\}$, $\{u_7, u_8, u_{10}\}$ sont des exemples de cliques, alors que l'ensemble $\{u_1, u_2, u_4, u_6\}$ n'est pas une clique, puisque u_2 et u_6 ne sont pas synonymes. De

10. Ce dictionnaire, ainsi qu'une grande partie des traitements que nous avons réalisés, seront bientôt accessibles par Internet, sur le site Web de l'ELSAP qui est en cours de réalisation (http://www.elsap.unicaen.fr).
11. Il faut noter que ce que nous appelons *clique* ici est appelé parfois *clique maximale* dans la littérature. Pour une présentation de la théorie des graphes, voir par exemple Bergé C. 1970, Gondran M. & Minoux M. 1975, ou encore Labelle J. 1981.

même {u_6, u_9} ne forme pas une clique parce qu'elle n'est pas maximale : ces deux unités sont toutes deux synonymes de u_{10} et c'est donc {u_6, u_9, u_{10}} qui constitue une clique. Comme on peut le constater, deux cliques peuvent être disjointes, ou avoir une ou plusieurs unités en commun.
Si l'on se centre maintenant sur une unité donnée, l'ensemble des cliques contenant cette unité constitue un recouvrement de l'ensemble des synonymes de cette unité : tout synonyme de l'unité appartient à au moins l'une de ces cliques, et réciproquement tout élément d'une de ces cliques est un synonyme de l'unité considérée. L'ensemble des cliques ainsi associé à une unité lexicale révèle donc la structure de l'ensemble des synonymes de cette unité. On peut en particulier traduire en termes de cliques les définitions que nous avons données de la monosémie, de l'homonymie et de la polysémie :
- à une unité monosémique correspond un ensemble de cliques réduit à une seule clique ; ainsi l'unité u_5 ne forme qu'une seule clique avec ses 3 synonymes : {u_5, u_2, u_4, u_7}. Elle est donc monosémique, comme d'ailleurs u_3 et u_8.
- à une unité homonymique correspond un ensemble de cliques que l'on appellera *séparable* ; c'est le cas de l'unité u_7, qui possède 6 synonymes et 3 cliques, qu'on peut séparer en deux sous-ensembles : d'une part { u_7, u_2, u_4, u_5} et d'autre part {u_7, u_8, u_{10}} et {u_7, u_9, u_{10}}, qui sont tels que chaque clique de l'un des sous-ensembles n'a pas d'élément en commun avec aucune clique de l'autre sous-ensemble, à part l'unité u_7 elle-même. On vérifiera que u_1 et u_6 sont aussi homonymiques dans notre exemple.
- à une unité polysémique correspond un ensemble de plusieurs cliques que l'on appellera *non séparable* ; par exemple l'unité u_4 possède 5 synonymes et 3 cliques, {u_4, u_1, u_2}, {u_4, u_1, u_6} et {u_4, u_2, u_5, u_7}, que l'on ne peut pas séparer (u_1 est commun à la première et à la deuxième, et u_2 à la première et à la troisième). De même, u_2, u_9 et u_{10} sont aussi polysémiques.
Comme on peut le constater, la structure de l'ensemble des cliques associé à une unité donnée est propre à cette unité : deux unités, même si elles sont synonymes, ne possèdent pas en général le même ensemble de cliques. Cette structure semble donc bien adaptée pour rendre compte de la spécificité sémantique de chaque unité lexicale. En fait, nous allons voir qu'elle permet effectivement de dégager les différents sens d'une unité, sans pour autant écarter les sens intermédiaires qui caractérisent les unités polysémiques.

2.3. Les cliques du graphe des synonymes

La recherche des cliques d'un graphe est très coûteuse en temps machine, si l'on utilise un algorithme non optimisé. Pour obtenir les cliques associées à une unité lexicale donnée à partir de notre dictionnaire des synonymes, nous avons implémenté la méthode proposée dans Reingold E.M. 1977, et Bron C. & Kerbosch J. 1973. Celle-ci permet de n'engendrer, en parcourant le graphe, qu'une seule occurrence de chaque clique[12].
Le nombre de cliques associées à une unité varie beaucoup selon l'unité lexicale considérée. Le tableau de la figure 2 en donne quelques exemples qui illustrent bien cette variabilité. Ces chiffres « bruts » sont encore sujets à révision. En effet, au-delà des erreurs typographiques et autres (cf. note 9) dont la correction ne pose pas de problème théorique, nous avons rencontré d'autres difficultés beaucoup plus délicates à traiter. Sans entrer dans les détails ici, signalons les variantes orthographiques, la présence de locutions, mots composés, expressions figées, etc., les phénomènes de polycatégorie grammaticale (nom-verbe, nom-adjectif, adjectif-adverbe, etc.). Pour limiter la prolifération incontrôlée de formes à laquelle peuvent conduire ces phénomènes, il faut faire des choix qui comportent inévitablement une certaine part d'arbitraire (cf. Fuchs C. *et al.* 1993 : 88-92). Nous n'avons pas encore pris de décision définitive sur ces questions. Plutôt que de trancher prématurément, il nous a semblé préférable de commencer par un travail exploratoire sur quelques

12. Jean-Luc Lambert (GREYC, Université de Caen) nous a aidés dans la recherche de cet algorithme optimal.

exemples, sans toucher au dictionnaire dans son ensemble. Notre premier objectif était de trouver d'abord le type de traitements qui permettraient d'exploiter au mieux les données.

Unité lexicale	Nombre de synonymes	Nombre de cliques
défendre	72	50
interdire	39	46
centre	42	39
maison	103	130
démarcation	9	7
insensible	73	94
faible	162	280
superbement	9	5

Figure 2 : Quelques exemples de nombres de synonymes et de cliques

La première constatation qui s'impose, quand on observe les cliques associées à un certain nombre d'unités lexicales, c'est que ces cliques semblent représenter un niveau très fin de granularité du sens. Certaines cliques représentent les sens les plus typiques des unités, et d'autres des sens intermédiaires, qui illustrent très précisément l'existence d'un continuum entre des sens typiques parfois très éloignés les uns des autres. Quelques exemples suffiront à en convaincre le lecteur.

Certaines cliques de *défendre* correspondent à des nuances de sens qui jouent sur la valeur « positive » que nous avions relevée ci-dessus (§1.3) :[13]

couvrir, excuser, justifier
disculper, excuser, justifier
plaider, soutenir
intercéder, intervenir
assurer, maintenir, protéger, soutenir
abriter, garantir, protéger, préserver
appuyer, maintenir, soutenir, tenir
abriter, couvrir, garantir, protéger
chaperonner, couvrir, garantir, protéger
garder, maintenir, tenir

D'autres correspondent à la valeur « négative » :

empêcher, inhiber, interdire, prohiber
inhiber, interdire, prohiber, proscrire
condamner, empêcher, interdire, prohiber
censurer, condamner, interdire, prohiber, proscrire
censurer, condamner, réprouver

Mais il n'y pas de coupure : l'ensemble des cliques de *défendre* n'est pas séparable. Toute une série de cliques montrent comment l'on passe progressivement de l'une à l'autre, démontrant ainsi l'unité du sémantisme de ce verbe :

excuser, justifier, sauver
garantir, garder, protéger, préserver, sauvegarder, sauver
disputer, soutenir

13. Dans tous les exemples de cliques qui suivent, nous avons systématiquement omis l'unité étudiée, qui fait bien sûr partie elle aussi de toutes les cliques qui lui sont associées (les autres unités de chaque clique sont présentées par ordre alphabétique).

intervenir, secourir
aider, appuyer, fortifier, protéger, renforcer
garder, maintenir, tenir
couvrir, flanquer, protéger, renforcer
empêcher, tenir
empêcher, interdire, s'opposer
refuser, s'opposer

Autre exemple, l'adjectif *insensible* : deux ensembles de sens s'organisent autour de deux constructions possibles de *sentir*, dont le sujet peut désigner le siège de la sensation (d'où *insensible* = « qui ne peut pas éprouver de sensation ») ou la source de la sensation (d'où *insensible* = « qui ne peut pas causer de sensation »). En fait, ces deux sens sont reliés par une série de cliques intermédiaires :

endormi, engourdi, indolent
engourdi, froid, inerte
frigide, froid, glacé
apathique, indifférent, indolent
flegmatique, froid, impassible, imperturbable, indifférent
dur, froid, inaccessible, indifférent
impénétrable, inaccessible, insaisissable, sourd
imperméable, impénétrable, inabordable, inaccessible, indifférent
imperceptible, indiscernable, insaisissable, invisible
indifférent, insignifiant, neutre
imperceptible, inapparent, invisible
insignifiant, léger, négligeable
imperceptible, insignifiant, léger

De plus, un certain nombre de cliques révèlent une autre gradation, qui porte sur le caractère plus ou moins volontaire de l'absence de sensation :

indifférent, sans-coeur, sec, égoïste
cruel, dur, féroce, impitoyable, implacable, inexorable, inhumain
dur, rigide, stoïque, sévère
dur, froid, glacial, sec
impassible, imperturbable, implacable, inflexible
imperméable, impénétrable, inaccessible, réfractaire, sourd
blasé, flegmatique, froid, indifférent
détaché, indifférent, étranger
calme, immobile, impassible
assoupi, endormi, engourdi
apathique, endormi, inerte
engourdi, immobile, inerte, paralysé
calme, immobile, inanimé
apathique, inerte, mort
froid, inanimé, inerte
inanimé, inerte, mort

Comme on peut le vérifier, les cliques déterminent des sens beaucoup plus précis et étroits que les synonymes eux-mêmes, qui englobent pour certains d'entre eux (*froid, indifférent,* etc.) une bonne partie de la polysémie de *insensible*.

Prenons, comme dernier exemple, le nom *maison*. Cette unité est fondamentalement polysémique[14]. On y trouve pourtant des sens très divers, comme le montrent les cliques suivantes :

boîte, entreprise, firme, établissement
branche, famille, lignée, race
baraque, bicoque, cabane, cahute, habitation
building, bâtiment, bâtisse, construction, immeuble, édifice
château, hôtel, palais
cabane, prison
demeure, domicile, habitation, logement, résidence, séjour
chez-soi, home, intérieur
famille, maisonnée, ménage
domesticité, domestique, serviteur

Il faut noter qu'un grand nombre de cliques décrivent avec une extraordinaire précision le type d'habitation que peut désigner le mot *maison*, comme le montrent ces quelques extraits :

bouge, galetas, réduit, taudis
bicoque, gourbi, masure, taudis
appartement, bouge, taudis
baraque, cabane, cahute, habitation, hutte
cabane, cahute, réduit
bicoque, cabane, maisonnette
cabane, case, hutte, maisonnette
cabane, chaumière, maisonnette
bâtisse, masure
chalet, habitation, pavillon, villa
maisonnette, pavillon
appartement, habitation, logement, logis
bâtiment, habitation, immeuble
building, bâtiment, bâtisse, construction, immeuble, édifice
abri, château, habitation, pavillon
château, demeure, habitation, palais
habitation, immeuble, palais
hôtel, immeuble, palais

Ainsi les cliques de synonymes caractérisent des sens très précis de l'unité étudiée, tout en ne masquant pas les relations de voisinage que ces sens entretiennent. Dans le cadre du modèle continu que nous avons présenté plus haut, les cliques semblent donc pouvoir représenter des régions très restreintes de l'espace sémantique, qui recouvrent à elles toutes l'espace tout entier. Elles devraient donc permettre de construire cet espace sémantique, si l'on trouve le moyen de tirer de la structure du graphe une relation de voisinage entre les régions associées aux cliques.

3. Construction automatique de l'espace sémantique

3.1. Une métrique pour l'espace des cliques

La théorie des graphes offre de nombreuses techniques qui permettent de mettre en évidence différentes caractéristiques de la structure d'un graphe. Elles sont donc a priori exploitables pour

14. Plus précisément, sur les 130 cliques initiales, si l'on ne tient pas compte de quelques cas douteux (comme *chez soi* sans trait d'union) et des deux seules locutions présentes (*train de vie* et *train de maison*), on obtient un sous-ensemble non séparable de 115 cliques, et moins d'une dizaine de synonymes isolés chacun dans une clique (c'est le cas de *couronne, mesnil, pigeonnier, standing, temple, trône*).

décrire les relations qu'entretiennent les cliques du graphe. Mais après un certain nombre de tentatives en ce sens, nous avons acquis la conviction[15] que la relation de voisinage entre cliques que nous recherchions ne pouvait véritablement être exprimée qu'en définissant une *distance* entre cliques. Autrement dit, il faut considérer que l'espace sémantique associé à une unité est un *espace métrique*, et que les petites régions correspondant aux cliques peuvent être assimilées à des *points* de cet espace. Les relations entre cliques sont alors représentées par les distances respectives entre ces points.

Nous avons donc cherché une représentation géométrique adaptée à notre problème. Le plus simple en principe consiste à se placer dans l'espace euclidien engendré par les synonymes, et à faire correspondre à chaque clique un sommet d'un hypercube de cet espace. En effet, si l'on appelle u_1, u_2, …, u_n les synonymes, et c_1, c_2, …, c_p les cliques associées à l'unité étudiée, le synonyme u_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 le synonyme correspondant appartient ou non à la clique :

$x_{ki} = 1$ si $u_i \in c_k$ et $x_{ki} = 0$ si $u_i \notin c_k$

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 :

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

Mais cette distance se révèle totalement inadéquate. L'analyse de quelques exemples, comme celui que nous allons présenter dans un instant, nous en a rapidement convaincus. Cela provient du 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 comme on l'a vu, certains synonymes peuvent recouvrir une grande partie des emplois de l'unité, 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. Il fallait donc trouver[16] une autre métrique répondant à ces besoins. C'est une métrique bien connue en analyse de données[17], *la métrique du* χ^2, qui s'est avérée satisfaisante. On peut la définir de la façon suivante :

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

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

Autrement dit, cette distance possède les deux caractéristiques suivantes. D'une part, chaque synonyme, en tant que vecteur de base de l'espace, intervient dans le calcul avec un « poids » plus faible si le synonyme est présent dans un grand nombre de cliques : les synonymes qui sont les moins spécifiques jouent un rôle moins important dans la discrimination des sens de l'unité. D'autre part, les coordonnées de chaque clique sont divisées par le nombre d'éléments de la clique : le point représentant la clique n'est donc plus un sommet de l'hypercube mais il est d'autant plus proche de l'origine que la clique correspondante comporte plus de synonymes.

---

15. L'ensemble de cette démarche est décrit dans Ploux S. *à paraître*. On y trouvera en particulier les principaux résultats obtenus par ces méthodes de théorie des graphes, ainsi qu'une discussion sur leur intérêt respectif.

16. Tous nos remerciements à Benoît Habert (ELI, ENS Fontenay-St Cloud) , Adeline Nazarenko (LIPN, Université Paris-Nord) et à Yves-Marie Visetti (LIP6, Université Paris 6), dont la contribution au cours d'un atelier de travail a été décisive à ce stade de la recherche.

17. Pour une introduction à toutes les notions d'analyse de données utilisées dans la suite de cet article, voir par exemple Bouroche J.-M. & Saporta G. 1994, Diday E. *et al.* 1982, Bry X. 1995 ou encore Volle M. 1997.

Pour donner une idée intuitive de ces définitions, considérons deux couples de cliques, $\{c_1, c_2\}$ et $\{c_3, c_4\}$, et supposons que dans chaque couple, les deux cliques diffèrent par un même nombre de synonymes. Pour la métrique canonique, ces deux couples sont analogues : la distance entre les cliques c_1 et c_2 est égale à la distance entre les cliques c_3 et c_4. En revanche, il n'en est pas forcément de même pour la métrique du χ^2. Si le premier couple $\{c_1, c_2\}$ est constitué de deux cliques nombreuses qui possèdent beaucoup de synonymes « spécifiques » communs, la distance entre les cliques c_1 et c_2 sera très faible. Et si le deuxième couple $\{c_3, c_4\}$est constitué au contraire de deux cliques peu nombreuses qui diffèrent par des synonymes spécifiques, la distance entre les cliques c_3 et c_4 sera nettement plus grande. Cela correspond beaucoup mieux à la conception que l'on peut se faire de l'écart plus ou moins grand des sens associés à ces cliques.

Pour être tout à fait concrets, nous allons prendre un exemple, issu de l'analyse du verbe *interdire*. Parmi ses nombreuses cliques, on en trouve beaucoup qui caractérisent des nuances de son sens usuel, comme les deux suivantes :

c_{26} : *défendre, inhiber, prohiber, proscrire*

c_{32} : *empêcher, inhiber, paralyser, supprimer*

Mais on trouve aussi des cliques qui correspondent à un sens bien différent, que l'on trouve surtout dans les emplois adjectivaux du participe passé, comme dans l'énoncé :

Devant tant de mauvaise foi, Jean est resté tout interdit.

Parmi les cliques qui caractérisent ce type d'emplois[18], on trouve :

c_{46} : *paralyser, pétrifier*

On s'aperçoit que les cliques c_{46} et c_{32} possèdent un synonyme en commun (*paralyser*, qui peut prendre lui aussi les deux sens) et diffèrent par 4 autres synonymes en tout, tandis que les cliques c_{32} et c_{26}, qui ont aussi un synonyme en commun (*inhiber*), diffèrent par pas moins de 6 autres synonymes. La distance canonique aboutit donc à une représentation aberrante, dans laquelle c_{32} est un peu plus proche de c_{46} que de c_{26} !

En revanche, la distance du χ^2 rectifie de façon spectaculaire cette aberration, comme le montrent les schémas de la figure 3.

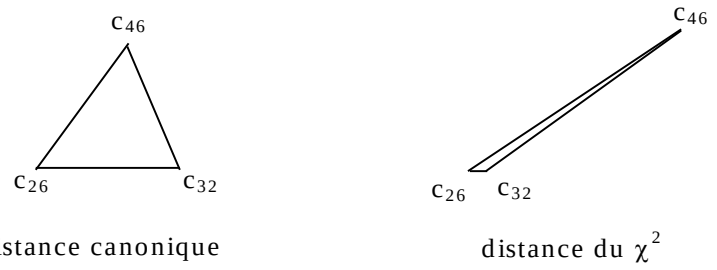

Figure 3 : Comparaison des deux distances

La distance du χ^2 confère donc à l'ensemble des cliques une structure géométrique qui semble respecter la notion intuitive de proximité entre sens d'une unité. La voie est ainsi ouverte à la construction automatique de l'espace sémantique associé à une unité, si l'on accepte d'identifier chaque clique à un point de cet espace sémantique. Il faut noter que si l'unité est monosémique, on

18. On peut se poser la question de savoir s'il s'agit vraiment d'emplois du verbe, ou plutôt d'un adjectif dérivé du verbe. Ainsi le *Petit Robert* possède une entrée *interdit (adj.)* qui relève (entre autres) ces emplois. Mais le même *Petit Robert* signale aussi à l'entrée *interdire* un sens (vieilli, certes) du verbe pour ces mêmes emplois (avec comme exemple ce vers de Regnard : *Et ce brusque discours a de quoi m'interdire*). Comme nous le disions plus haut, ces questions de polycatégorie sont difficiles à trancher. On y reviendra à propos de cet exemple précis au §3.3, note 22.

n'a qu'une seule clique, ce qui signifie que cet espace se réduit à un point ; alors que si l'unité est homonymique, l'ensemble des cliques est séparable, et l'espace sémantique peut être considéré comme l'union de plusieurs sous-espaces que l'on peut étudier séparément. Cela correspond parfaitement aux attendus de notre modèle théorique.
Le problème, c'est que l'ensemble des cliques, dans la représentation que nous avons choisie, est « plongé » dans un espace de très grande dimension (égale au nombre de synonymes de l'unité étudiée). Il est de ce fait peu maniable[19], et surtout cela masque une propriété importante que nous avons postulée pour l'espace sémantique : son petit nombre de dimensions (cf. §1.3). Si cette hypothèse est exacte, l'ensemble des cliques n'occupe pas de manière plus ou moins homogène l'espace engendré par les synonymes. Il est au contraire confiné à une région que l'on doit pouvoir décrire comme une approximation d'un sous-espace de faible dimension.
Pour en juger, on dispose d'une méthode d'analyse des données, *l'analyse en composantes principales*[20], dont c'est très précisément la fonction. En effet cette méthode consiste, en gros, à déterminer une suite d'axes orthogonaux, centrés sur le centre de gravité d'un nuage de points, tels que la projection du nuage de points sur l'espace engendré par ces axes soit la moins « déformée » possible (en un sens précis). Ces axes sont ordonnés par importance décroissante : si un petit nombre d'axes suffit à rendre compte de l'essentiel de la « dispersion » de cet ensemble de points, on peut alors considérer avec une bonne approximation que ces points se situent tous sur le petit sous-espace engendré par ces axes. Dans notre cas, cela signifie que si un petit nombre d'axes suffit pour représenter les cliques associées à une unité, l'espace sémantique associé à cette unité est approximativement contenu dans le sous-espace correspondant.
Nous avons donc utilisé cette méthode pour étudier la possibilité de construire automatiquement l'espace sémantique associé à une unité lexicale. Les premiers résultats que nous avons obtenus sont très encourageants. Pour les illustrer, nous allons présenter et commenter deux exemples particulièrement significatifs.

3.2. L'exemple du nom *maison*

Si l'on applique l'ensemble des traitements que nous venons de décrire à l'unité lexicale *maison*, on obtient, si l'on se limite aux deux premiers axes de l'analyse en composantes principales, la représentation illustrée figure 4. Chaque point représente une clique, et nous avons fait figurer le contenu de certaines d'entre elles.
Comme on peut le constater, les cliques se distribuent le long d'une courbe, allant de sens de type « établissement commercial », jusqu'à des sens exprimant la notion de descendance, en passant par une valeur centrale de lieu d'habitation, représentée par une majorité de cliques. Cette structure présente plusieurs caractéristiques dignes d'intérêt.

19. Encore que certaines méthodes d'analyse des données (comme par exemple la classification par nuées dynamiques) pourraient être utilisées directement dans cet espace de grande dimension.
20. L'analyse en composantes principales des lignes (ou des colonnes) d'un tableau de données qualitatives, munies de la métrique du χ^2, fait partie de ce que l'on appelle *l'analyse factorielle des correspondances* (cf. Bouroche J.-M. & Saporta G. 1994 : 92-96).

Figure 4 : Représentation des cliques associées à *maison*

D'abord, les deux axes sont susceptibles de recevoir une interprétation sémantique. Le premier, vertical, semble s'organiser de haut en bas selon un gradient qui va de la « facette » socio-économique du mot *maison* (*La maison ne fait pas crédit*, *Tous nos produits sont fait maison*) à sa facette socio-anthropologique (*La maison de Habsbourg*, *La maison des Bourbon*), les valeurs médianes étant occupées, comme de juste, par la maison comme habitation et lieu de vie de ce que l'on peut appeler « le noyau familial », constitué d'individus apparentés et vivant ensemble, qui constitue bien l'unité de base, à la fois économique et anthropologique, de nos sociétés. Le deuxième, horizontal, va plutôt des emplois de maison plus « concrets », désignant les entités physiques que sont les bâtiments d'habitation jusqu'à des emplois plus « abstraits », désignant des institutions humaines dues à des relations sociales, qu'il s'agisse de rapports commerciaux ou de liens de descendance.

Il faut aussi noter que les différents types d'habitation se retrouvent groupés au centre de la courbe. La prise en compte d'un troisième axe aurait peut-être permis de rendre compte de la diversité de ces emplois, dont nous avons vu plus haut (§2.3) qu'elle était considérable. Mais en fait, on peut

aussi se satisfaire de cette absence de discrimination : après tout *maison* est justement le terme générique qui englobe tous les types d'habitation, et leur différenciation n'est pas du ressort du sémantisme de cette unité.
Enfin, il faut souligner tout l'intérêt d'avoir obtenu une représentation qui épouse approximativement la forme d'une courbe. Cela signifie en effet que si l'on ne cherche pas de finesse excessive, on peut parcourir l'ensemble des sens de *maison* à l'aide d'un seul paramètre. Autrement dit l'espace sémantique associé à *maison* est en première approximation un espace unidimensionnel (variété de dimension 1). C'est très important si l'on veut se servir de cet espace pour calculer le sens de *maison* dans un énoncé donné, puisque cela montre qu'une approche continuiste de la polysémie, loin de multiplier à l'infini les paramètres nécessaires pour représenter le sens, permet au contraire de *simplifier* la représentation de manière radicale. Ainsi, si l'on se place dans le cadre du modèle théorique sur lequel nous nous appuyons (cf. §1.3), le sens de *maison* dans n'importe quel énoncé pourra être déterminé par un intervalle (ou plusieurs, en cas d'ambiguïté-alternative) de valeurs de l'unique paramètre, ou si l'on préfère, par un « morceau » de la courbe représentative de l'espace sémantique. Cette simplicité pourrait donc être mise à profit dans une réalisation informatique pour simuler un calcul dynamique du sens (qu'on peut implémenter par exemple à l'aide d'un réseau connexionniste : cf. Victorri B. & Fuchs C. 1996).

3.3. L'exemple du verbe *interdire*

Le deuxième exemple, le verbe *interdire*, va nous permettre d'illustrer d'autres propriétés des représentations auxquelles nous avons abouti. Si l'on se limite là encore à deux axes principaux, on obtient pour *interdire* le schéma présenté figure 5.
On constate immédiatement que les cliques correspondant aux emplois adjectivaux de *interdit* indiquant la stupéfaction (cf. plus haut, fin du §3.1) se retrouvent confinées dans une toute petite région, bien séparée des autres cliques. Cela est d'autant plus remarquable que ces emplois sont très riches en synonymes.
On voit aussi à l'œuvre une autre des possibilités qu'offre cette représentation : on peut y faire figurer les synonymes eux-mêmes, en les plaçant au centre de gravité des cliques qui les contiennent. Cela permet de comprendre d'un seul coup d'œil pourquoi l'ensemble des cliques n'est pas séparable, malgré cette coupure visible sur la figure. En effet, on constate que *paralyser* et, moins nettement, *saisir* se situent entre les deux groupes de cliques : ce sont des éléments pivots, qui participent aux deux familles de cliques[21], et qui empêchent *interdire* d'être homonymique.

21. Les cliques auxquelles participe *saisir* sont les suivantes : {*saisir, supprimer*}, {*saisir, troubler, étonner*} et {*pétrifier, saisir, étonner*}. Pour *paralyser*, voir la fin du §3.1.

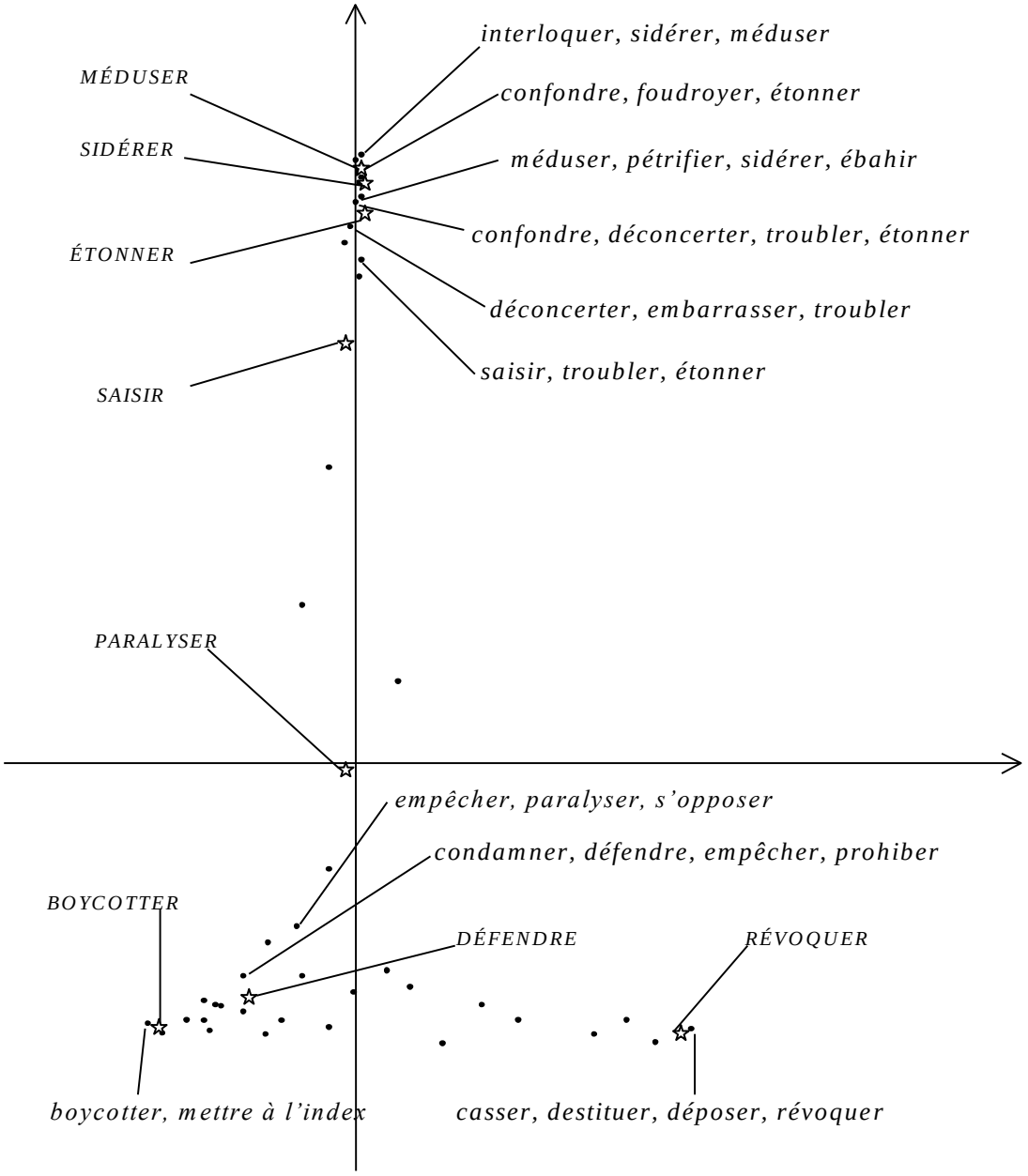

Figure 5 : Représentation de *interdire*.
Les points représentent les cliques et les étoiles quelques uns des synonymes (mots en majuscule), placés au centre de gravité des cliques qui les contiennent.

On a donc tous les éléments en main pour prendre la décision qui convient : par exemple, dans un cas comme celui-ci, il peut sembler raisonnable de décider, malgré l'existence d'éléments pivots, de traiter *interdire* comme deux homonymes, $interdire_1$ et $interdire_2$, en opérant une partition de l'ensemble des cliques. Chacun des deux sous-ensembles peut alors faire l'objet, indépendamment l'un de l'autre, d'une nouvelle analyse, ce qui permettrait de mieux représenter les deux espaces sémantiques disjoints ainsi postulés. On peut d'ailleurs espérer que cette décision se prenne automatiquement (en utilisant un algorithme de classification avec un seuil élevé : la partition n'est effectuée que si l'inertie interclasse est très grande relativement à l'inertie intraclasse). L'approche continue n'est donc nullement incompatible avec une prise en compte des cas d'homonymie, même

« masqués ». Au contraire, la méthode proposée semble permettre d'identifier ces cas de manière simple[22].

Conclusion

Ainsi ce premier travail exploratoire semble tout à fait prometteur, même si, bien entendu, les quelques exemples que nous avons traités jusqu'à présent ne suffisent pas. Il nous faut encore effectuer un gros travail de validation, en analysant les résultats pour un nombre important d'unités lexicales très diverses, avant de pouvoir conclure définitivement. Mais cela montre déjà que l'approche continuiste de la polysémie lexicale se prête bien à des traitements automatiques qui utilisent les ressources offertes par les dictionnaires électroniques.

On peut d'ailleurs imaginer toutes sortes d'extensions au travail qui a été présenté ici. Par exemple, on s'est limité ici à la construction de l'espace sémantique associé à *une* unité lexicale. Or on peut tout aussi bien appliquer les mêmes traitements à l'ensemble des cliques associées à *plusieurs* unités lexicales à la fois : bien que cela n'ait pas été très apparent dans notre présentation, la définition de la distance entre cliques ne fait pas jouer de rôle particulier à l'unité que l'on a choisi d'étudier. On peut donc imaginer de traiter tout un paradigme lexical, en construisant un espace sémantique associé à ce paradigme, qui serait en quelque sorte une représentation du « recollement » des espaces sémantiques de chacune des unités du paradigme. Une telle représentation permettrait de mieux comprendre comment ces unités peuvent se définir les unes par rapport aux autres, dans quelle mesure elles s'opposent, elles se recouvrent, etc.

Dans le même ordre d'idées, on peut aussi envisager de travailler avec des dictionnaires bilingues (peut-être même fusionnés avec des dictionnaires de synonymes de chacune des langues), et de représenter dans un même espace les cliques « mixtes » composées d'unités lexicales des deux langues. Au plan des applications, cela permettrait peut-être de mettre au point des outils efficaces pour l'aide à la traduction. Au plan théorique, cela peut aussi se révéler intéressant pour étudier comment chacune des langues « découpe » un même champ sémantique.

Enfin à plus long terme, on peut imaginer de coupler ce type de travail avec des analyses de corpus, de façon à combiner les propriétés des unités lexicales à la fois sur l'axe paradigmatique et sur l'axe syntagmatique. Il semble possible en effet d'utiliser un modèle continu pour représenter le comportement des unité lexicales à la fois sur ces deux axes (cf. la tentative de Piotrowski D. 1997 dans cette direction).

Références

BERGE C. (1970) : *Graphes et hypergraphes*, Paris, Dunod.

BOUROCHE J.-M. & SAPORTA G. (1994) : *L'analyse des données*, Que sais-je ?, Paris, PUF, 6ème édition.

BRON C., KERBOSCH C. (1973) : « Algorithm 457 : Finding All Cliques of an Undirected Graph », *Communications of the ACM*, n° 16, 575-577.

BRY X. (1995) : *Analyses factorielles simples*, Paris, Economica.

DIDAY E., LEMAIRE J., POUGET J., TESTU F. (1982) : *Eléments d'analyse des données*, Paris, Dunod.

22. Pour en revenir à notre discussion sur le problème de la polycatégorie (cf. §3.1 note 18), ce genre de résultats laisse à penser qu'il est plus judicieux de ne pas décider a priori de séparer les sens des unités polycatégorielles : en fait, la méthode que nous proposons semble capable de réaliser automatiquement, quand il y a lieu, ces séparations. Il reste bien sûr à vérifier si ce joli résultat est bien représentatif du cas général…

DUTOIT D. (1992) : « A comprehensive approach to lexical semantics », *COLING'92*, Nantes, 982-987.

FUCHS C., DANLOS L., LACHERET A., LUZZATI D., VICTORRI B. (1993) : *Linguistique et traitement automatique des langues*, Paris, Hachette.

GONDRAN M. & MINOUX M. (1979) : *Graphes et algorithmes*, Paris, Eyrolles.

GREFENSTETTE G. (1994) : *Explorations in Automatic Thesaurus Discovery*, Dordrecht, Kluwer.

HABERT B., NAZARENKO A., SALEM A. (1997) : *Les linguistiques de corpus*, Paris, Armand Colin.

HINDLE D. (1990) : « Noun classification from predicate-argument structures », *ACL'83*, Berkeley, 268-275.

LABELLE J. (1981) :*Théorie des graphes*, Montréal, Modulo.

MILLER G.A., BECKWITH R., FELLBAUM C., GROSS D., MILLER K. (1993) : *Five Papers on WordNet*, http://www.cogsci.princeton.edu/wn/.

PIOTROWSKI D. (1997) : *Dynamiques et structures en langue*, Paris, CNRS Editions.

PLOUX S. (*à paraître*) : « Modélisation et traitement informatique de la synonymie », *Linguisticæ Investigationnes*.

REINGOLD E.M., NIEVERGELT J., DEO N. (1977) : *Combinatorial Algorithms, Theory and Practice*, Prentice-Hall.

VÉRONIS J., IDE N. (1990) : « Word sense disambiguation with very large neural networks extracted from machine-readable dictionaries », *COLING'90*, Helsinki, 389-394.

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

VOLLE M. (1997) : *Analyse des données*, Paris, Economica.

WARNESSON I. (1985) : « Applied Linguistics : Optimization of Semantic Relations by Data Aggregation Techniques », *Journal of Applied Stochastic Models and Data Analysis*, Vol. 1, n°2, 121-143.

WARNESSON I. (1992) : « Lexicographie et informatique, vers une nouvelle génération de dictionnaires », *Publications scientifiques et techniques d'IBM France,* décembre 1992, Paris, 107-157.