# La représentation de la sémantique des phrases dans le paradigme géométrique des Atlas sémantiques : Une étude sur les syntagmes de type V-N et N-ADJ

Sabine Ploux
L2C2, Institut des Sciences Cognitives-CNRS, Université Lyon I, Bron, France

**Abstract** This article explores the ability of the geometric paradigm of the Semantic Atlases (SA) to represent the semantics of verb and noun phrases. The SA model assigns each lexical unit an area in a multidimensional space. The area contains the set of semantic values of a lexical unit and depicts their distances from each other. In order to study the semantics of phrases, we defined a composition operator that uses this geometric representation. The operator positions a given phrase in the space of phrases closest to it in meaning, and determines the specific value taken on by each word in the phrase. This theoretical framework was used to develop a preliminary computerized system. The results presented here describe the model's ability to simulate context-related variability of word meaning and the creation of meaning in new contexts.

**Résumé** Dans cet article, nous proposons d'explorer les capacités du paradigme géométrique des Atlas sémantiques (AS) à l'étude de la sémantique des syntagmes verbaux et nominaux. Les AS sont un modèle qui associe à chaque unité lexicale une forme dans un espace multidimensionnel. Cette forme représente l'ensemble des valeurs sémantiques de l'unité et figure leurs proximités relatives. Afin d'étudier la sémantique des expressions, nous avons défini un opérateur de composition qui s'appuie sur cette représentation géométrique. Cet opérateur permet de positionner une expression dans un espace d'expressions les plus proches du point de vue du sens et de déterminer la valeur spécifique prise par chacun des mots de l'expression. Le cadre théorique développé nous a permis de réaliser une première version d'un logiciel. Les résultats ici présentés rendent compte de la capacité du modèle à simuler la variabilité du sens d'un mot en contexte et la créativité du sens dans des contextes inédits.

# 1 Le sens des phrases et sa modélisation

## 1.1 Introduction

La description des mécanismes qui permettent de combiner des mots afin d'obtenir une phrase qui ait un sens est un thème de recherche de longue date ([Godart-Wendling et al., 1998] en donne un parcours historique). Le principe de compositionnalité sémantique, issu des travaux de Frege [Geach and Black, 1960], stipule que la signification d'une expression complexe est fonction de la signification de ses constituants et des règles syntaxiques en vertu desquelles elles sont combinées. Ce principe a motivé de nombreuses propositions dont la théorie sémantique de Montague [Dowty et al., 1981] qui utilise le cadre formel de la logique mathématique afin de définir les opérations de composition. Ce modèle logique ne prend pas pour support une théorie linguistique de la sémantique lexicale : les noms, par exemple, y ont pour valeur l'objet qu'ils désignent dans le monde. Mais plus récemment, plusieurs auteurs [Pustejovsky, 1998, Kintsch, 2001, Jackendoff, 1999] se sont intéressés à la modélisation des opérations sémantiques à l'œuvre dans la combinaison des constituants de la phrase. Les phrases traitées sont essentiellement les groupes nominaux constitués d'un adjectif et d'un nom ou encore les groupes verbaux constitués d'un verbe et de ses arguments. Par exemple, le verbe *bake* [Pustejovsky, 1998] dans *bake a cake* ou *bake the potatoes* désigne des processus différents. Dans le premier cas, il y a création d'un objet (*cake*); dans le deuxième cas, il ne s'agit que d'un changement d'état : les pommes de terre préexistent à la cuisson. De même, les adjectifs *good* et *fast* ont un sens chaque fois calculé en fonction des noms sur lesquels ils portent. Ainsi *un bon avocat* est un avocat qui défend bien ses clients, *un bon repas* est un repas savoureux, etc. On pourra se reporter à [Pustejovsky, 1998] pour une description plus détaillée des variations sémantiques de ces deux adjectifs.

Bien que, comme nous venons de le voir, le sens d'un mot diffère en fonction de son contexte d'emploi, il n'en reste pas moins que, sauf en cas d'homonymie dans lequel les sens sont séparés, les différentes occurrences ont des recouvrements sémantiques qui souvent s'organisent autour d'un noyau unique de sens. En somme, dans une expression le potentiel sémantique de chaque unité lexicale participe de façon différentielle et en fonction des autres éléments lexicaux et syntaxiques au sens global de l'expression. Et en retour, le sens de l'unité reçoit une valeur spécifique.

## 1.2 Des modèles en sémantique lexicale

Dans l'approche qui prend support sur la sémantique des constituants, la composition sémantique a donc pour préalable un modèle du lexique. Plusieurs propositions majeures ont été apportées dans ce domaine. La très connue base WordNet propose une organisation hiérarchique du sens lexical. Dans ce modèle, chaque lexème est décomposé en une liste de concepts lexicaux. Un concept est lui même représenté par un ensemble de synonymes (Synset) et étiqueté par

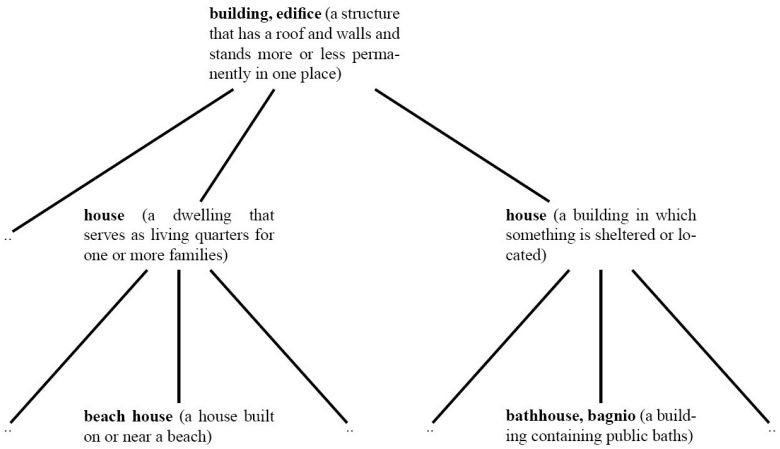


Figure 1: Extrait d'un sous-arbre de WordNet associé au concept *house* ; Dans cette figure, nous avons reproduit certains concepts lexicaux (en gras) et les propriétés qui leur sont associées telles quelles figurent dans la base (entre parenthèses).

un ensemble de propriétés qui le définissent. Les concepts lexicaux sont liés, au sein d'un graphe, par des relations lexicales (hyperonymie pour les noms, troponymie pour les verbes, etc.). La figure 1 donne une partie du graphe issu du mot *house*. La base WordNet offre l'avantage d'une grande couverture lexicale. Cependant, son architecture hiérarchique est plus adaptée aux noms qu'aux verbes, aux adjectifs et aux adverbes et enfin, sa constitution a été largement manuelle.

Les modèles vectoriels comme LSA [Landauer et al., 1998] ou HAL [Burgess and Lund, 1997] associent à un mot un vecteur dans un espace multidimensionnel. L'approche, basée sur une analyse statistique de corpus, offre l'avantage d'être totalement automatique. La figure 2 donne un aperçu d'une représentation vectorielle. Dans LSA, les proximités sémantiques entre mots sont calculées comme le cosinus de l'angle des vecteurs qui leur sont associés. Plus récemment, Kintsch [Kintsch, 2001] a étendu le modèle lexical en proposant un algorithme qui calcule les vecteurs associés non plus seulement aux mots mais aussi aux phrases de type N-V (exemple : *The horse ran*). Dans cette nouvelle proposition, le vecteur associé à une phrase est le barycentre des vecteurs associés au nom, au prédicat et à des termes voisins du prédicat et proches aussi du nom. Pour interpréter le sens de ces vecteurs, l'auteur choisit des mots discriminants (marqueurs). Ainsi, dans les phrases *The horse ran, vs. The color ran*, *gallop* (choisi comme marqueur) est plus proche de *The horse ran* que de *The color ran* alors que *dissolve* (choisi parce qu'il est proposé dans WordNet comme synonyme pertinent de *run* dans le contexte de *color*) est plus

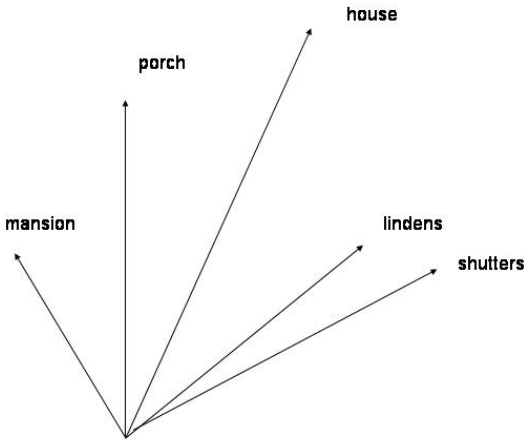


Figure 2: Figuration d'un mot et ses voisins dans un modèle vectoriel.

proche de *The color ran* que de *The horse ran*. En somme, dans ce modèle, le sens d'une expression ou celui des mots n'est pas caractérisé par une liste arrêtée de synonymes ou de propriétés mais interprété en fonction de la proximité à d'autres marqueurs. De plus, ces marqueurs sont choisis à la main et non directement donnés par le modèle.

Les approches hiérarchiques et vectorielles diffèrent par leurs hypothèses initiales : WordNet présuppose une organisation *a priori* du sens des mots, LSA part de l'hypothèse que le sens des mots ne peut être défini mais seulement mesuré par les ressemblances mutuelles qu'ils entretiennent entre eux. Cependant, ces deux approches ont un point commun : la représentation du sens par des unités atomiques insécables : un nœud dans un graphe ou un vecteur dans un espace. Pour cette raison, dans le cadre vectoriel, les différentes valeurs sémantiques d'un mot ne peuvent pas être représentées pour elles-mêmes à partir de l'objet mathématique (un vecteur) qui leur est associé à ce mot. « *Different meanings of the word or different senses of a word are not distinguished* », [Kintsch, 2001] . Dans Wordnet, ces différents sens sont fixés une fois pour toutes, leur granularité est celle du concept lexical (Synset). Les modèles WordNet et LSA sont essentiellement des modèles du lexique (même s'il existe comme nous venons de le voir des prolongements aux phrases). Pour rendre compte de la sémantique des combinaisons de mots, le modèle le plus élaboré reste certainement le Lexique génératif (LG) de Pustejovsky [Pustejovsky, 1998]. Pustejovsky fait le constat que pour rendre compte du sens des mots en contexte, on ne peut s'appuyer sur un lexique énumératif, c'est à dire un lexique pour

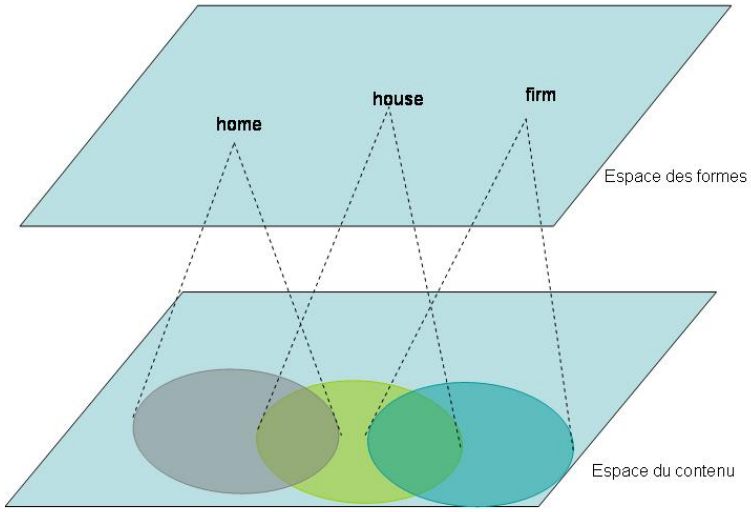


Figure 3: Figuration du lien entre mots et contenu sémantique dans un modèle géométrique.

lequel l'ensemble des sens associés à un mot est fixé et fini (comme c'est le cas par exemple dans WordNet). Dans le LG, chaque unité lexicale est dotée d'une structure interne d'attributs-valeurs en quatre niveaux : argumental (spécification du nombre et du type des arguments), événementiel (état, procès ou transition), de qualia (attributs du mot comme sa fonction ou son origine), d'héritage (position du mot dans un réseau lexical). Un ensemble de mécanismes génératifs permet de modéliser la formation d'unités composées et d'expliquer la modulation du sens en fonction du contexte. Cependant, Pustejovsky ne donne pas de méthodologie précise pour remplir les valeurs de chacun des attributs pour l'ensemble du lexique. Cette remarque rejoint celles de Kintsch [Kintsch, 2001] qui à la suite de Wittgenstein met en cause la possibilité même d'y parvenir.

Le paradigme géométrique dont nous rappelons maintenant les caractéristiques est une approche qui permet à la fois de rendre compte de la structure interne de la sémantique des mots (et ceci de façon automatique) sans pour autant nécessiter un métalangage de primitives dont la définition pose problème.

### 1.2.1 Qu'est ce qu'un modèle géométrique?

Nous appelons géométrique un modèle qui associe à un mot non plus un vecteur mais un domaine[1] dans un espace multidimensionnel. Ce choix permet à la fois

[1] Un domaine est tout d'abord un espace topologique, mais on suppose, de surcroît, que chacun de ses points possède un voisinage homéomorphe à un ouvert de $R^n$ (c'est à dire qu'il

(i) de représenter la structure interne de la sémantique d'un mot par interprétation des différentes zones qui constituent le domaine, (ii) de rendre compte de la similarité sémantique entre plusieurs unités lexicales à la fois par la mesure du recouvrement entre les domaines qui leur sont associés et par la distance qui les sépare. Nous faisons l'hypothèse de l'existence de deux niveaux celui des formes phonologiques et/ou orthographiques et un niveau substrat du contenu. Le domaine associé à un mot est alors la projection d'une forme (ici on considère les formes orthographiques) sur l'espace des contenus cognitifs. La figure 3 donne une idée de cette projection. Ainsi, les figures 1, 2, 3 résument, du point de vue de leur implémentation formelle, les différences entre trois grands types de modèles (hiérarchique, vectoriel, géométrique).

Le fait de distinguer ces deux niveaux permet à la fois de représenter les recouvrements sémantiques entre des lexèmes et les différentes valeurs sémantiques d'une unité lexicale mais aussi de rendre compte des différences entre langues [Ploux and Ji, 2003]. En effet, les langues n'opèrent pas les mêmes découpages extralinguistiques. On trouve des exemples de ces différences entre langues dans la désignation des couleurs. Chuquet et Paillard [Chuquet and Paillard, 1987] donnent également des exemples de ces différences de découpage entre le français et l'anglais : *room: pièce, chambre, bureau*, (ou dans un domaine abstrait) *esprit: mind, spirit, wit*. De plus, plusieurs études [Illes and Francis, 1999, Ikeda, 1998] ont permis de mettre en évidence un traitement et un système communs à la sémantique des langues chez des sujets bilingues. Ce partage d'un contenu cognitif indépendant des langues est figuré par le schéma de la figure 4.

Cependant, nous n'avons pas accès directement au contenu cognitif. Nous ne pouvons nous en approcher qu'à travers les traces que ce contenu induit sur les réalisations langagières. La notion de clique est un artefact utile pour nous en approcher. Nous allons rappeler maintenant les principes du modèle des AS et plus particulièrement la définition et les propriétés des cliques. Ces rappels nous serviront à présenter le modèle de composition de la sémantique des mots dans une expression.

# 2 Rappels

## 2.1 La sémantique des mots dans les AS

Nous utilisons l'exemple du mot *insensible* pour ces rappels.[2] Il permet de faire la synthèse des différentes caractéristiques du modèle. Le modèle des AS repose sur deux composantes : une base lexicale, le modèle géométrique à proprement

existe une bijection bi-continue du voisinage vers un ouvert de $R^n$). Par exemple, les aires délimitées par un disque, un carré, un rectangle ou une ellipse sont des domaines. Bien qu'un domaine soit lui-même un espace, dans ce texte, le terme domaine est plus spécifiquement utilisé pour désigner la forme géométrique associée à un mot, le terme espace est réservé à l'espace multidimensionnel $R^n$ qui le contient.

[2] Pour plus de détails, on pourra aussi se reporter aux publications précédentes [Ploux, 1997, Ploux and Victorri, 1998, Ploux and Ji, 2003] et tester les résultats du modèle sur le site des Atlas Sémantiques (http://dico.isc.cnrs.fr).

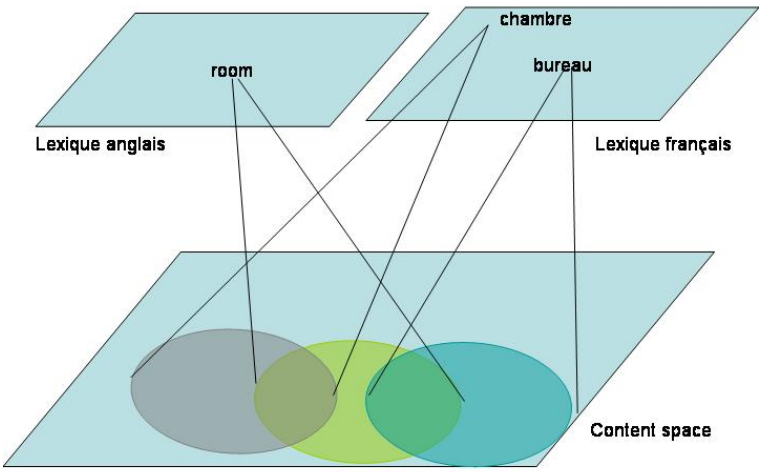


Figure 4: Figuration des projections de deux systèmes linguistiques sur un espace de contenu commun.

parlé.

### 2.1.1 Organisation de la base lexicale

WordNet définit les concepts lexicaux à partir de synonymes. En effet, un lien de synonymie entre deux mots marque le partage d'un contenu conceptuel. Dans un modèle géométrique, cette relation est figurée par un recouvrement entre les aires associées (voir figure 3). C'est aussi à partir d'une base de liens synonymiques et parasynonymiques que nous avons bâti le premier modèle géométrique de représentation de la sémantique des mots. Plusieurs bases lexicales (pour le français, l'anglais et pour l'espagnol) ont été crées (voir [Ploux, 1997] et [Ploux and Ji, 2003] pour le mode de création). La base du français, qui nous a aussi servi pour le modèle de composition ici présenté, contient 54,690 entrées. Chaque ligne d'une base lexicale est de la forme :

vedette : $similaire_1, similaire_2, similaire_3$...

Par exemple, pour la vedette *insensible*, la base contient 71 synonymes ou parasynonymes dont *anesthésié, apathique, aride, assoupi, blasé, calleux, calme, cruel, dur, détaché, endormi, endurci, engourdi, flegmatique, imperceptible, inapparent, inerte, sans-cœur...* Certains de ces mots représentent une valeur morale (*dur, sans-cœur...*) d'autres une valeur physique (*inerte, engourdi...*) d'autres enfin une valeur perceptive (*imperceptible, inapparent...*). Comment à partir de la liste initiale des synonymes distinguer et organiser ces différentes

valeurs sémantiques?

### 2.1.2 La notion de clique

Si la représentation associée à la sémantique d'un mot est, comme nous l'avons supposé, un domaine dans un espace multidimensionnel, nous devons, pour construire ce domaine, représenter les unités qui le composent. Ces unités qui sont de granularité plus fine que celle du mot lui même sont instanciées dans le modèle par des cliques. Les cliques sont des ensembles de mots tous synonymes les uns des autres. Si on considère le graphe de synonymie ayant pour nœuds les mots et pour arcs les liens attestés de synonymie entre deux mots, alors du point de vue de ce graphe, une clique est un sous graphe maximal complet connexe (voir figure 5 et [Ploux, 1997] pour une description détaillée). Du point de vue des domaines, c'est une intersection de plusieurs domaines associés à des un ensemble de synonymes tous synonymes les uns les autres (voir la figure 6). Une implication de cette définition est qu'il n'existe aucun autre mot dans la base lexicale qui puisse diviser l'intersection des domaines associés à la liste des mots de la clique. Pour cette raison, une clique représente une unité minimale de sens, un «grain» de sens. Cette granularité est plus fine que celle des Synsets de WordNet. À titre d'exemples, il existe 14 Synsets associés au mot *house* et 129 cliques pour le mot *maison* de même il existe 4 Synsets pour le mot *insensible* et 93 cliques pour le même mot en français. De plus cette granularité est indépendante des langues (voir [Ploux and Ji, 2003]).

Voici des exemples de cliques[3] qui figurent la valeur morale du mot *insensible* :

- 20 cruel, dur, impitoyable, implacable, inexorable, inflexible, inhumain, insensible
- 21 cruel, dur, impitoyable, implacable, inexorable, inflexible, insensible, sévère
- 22 cruel, dur, implacable, inflexible, inhumain, insensible, rigide
- 23 cruel, dur, implacable, inflexible, insensible, rigide, sévère
- ...

des exemples de cliques qui figurent la valeur physique :

- 2 anesthésié, insensible
- 50 endormi, engourdi, inerte, insensible
- 51 engourdi, froid, inerte, insensible
- 52 engourdi, immobile, inerte, insensible, paralysé

[3]Les cliques sont numérotées par ordre alphabétique telles qu'elles sont calculées par le modèle et figurent sur le site des AS.

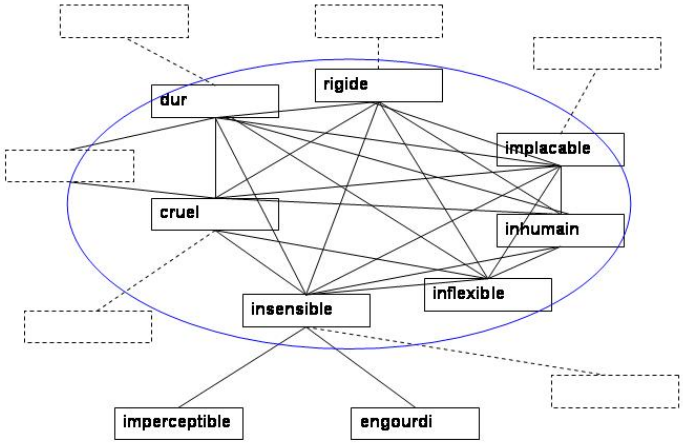


Figure 5: Une clique vue comme un sous graphe maximal complet connexe du graphe de synonymie.

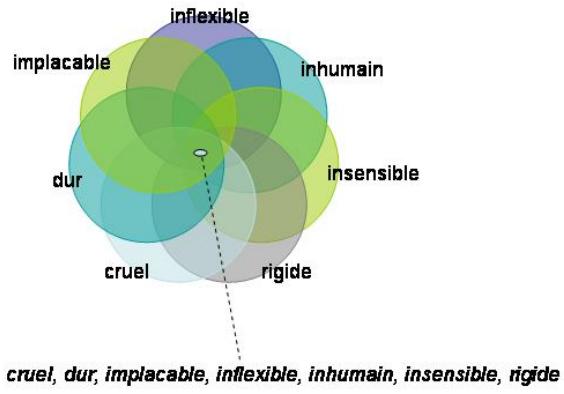


Figure 6: Une clique vue comme l'intersection des aires associées aux mots qu'elle comporte.

- ...

et enfin, des exemples pour la valeur perceptive :

- 69 imperceptible, inapparent, insensible, invisible
- 70 imperceptible, indiscernable, insaisissable, insensible, invisible
- 71 imperceptible, indiscernable, insensible, léger
- ...

On peut noter qu'un mot donné appartient à différentes cliques (cette caractéristique est due à la non transitivité de la relation de synonymie). Il apparaît dans chaque clique avec un sens précis qui est contraint par la présence de ses voisins. Ce partage des mots se retrouve, par exemple, dans le chemin

- 21 cruel, dur, impitoyable, implacable, inexorable, inflexible, insensible, sévère
- 34 dur, froid, impitoyable, implacable, insensible, sévère
- 35 dur, froid, inaccessible, indifférent, insensible
- 39 dur, impassible, indifférent, insensible, stoïque
- 15 calme, flegmatique, froid, impassible, imperturbable, insensible
- 16 calme, froid, inanimé, insensible
- 63 froid, inanimé, inerte, insensible
- 83 inanimé, inerte, insensible, mort

formé de cliques (dont chacune partage avec la suite au moins un terme) et qui fait passer d'une valeur morale à une valeur physique de façon relativement continue Cette topologie sous-jacente est mise en évidence dans la construction du domaine associé au mot initial, ici *insensible*.

### 2.1.3 Construction du domaine associé à un mot

Afin de construire le domaine géométrique, nous avons utilisé une méthode classique d'analyse factorielle des correspondances [Benzécri, 1980] entre les cliques et les termes synonymes. Cette analyse donne les coordonnées des cliques dans un espace multidimensionnel. Dans l'espace sémantique ainsi créé les cliques sont représentés par des points et chaque mot (mot vedette ou synonyme) par la région de l'espace délimitée par le nuage de cliques le contenant. Les écarts induits par les coordonnées représentent de façon cohérente les variations sémantiques (voir [Ploux, 1997, Ploux and Ji, 2003]). Enfin, un algorithme de classification sépare les différentes valeurs sémantiques associé au mot vedette. Pour le mot *insensible*, la figure 7 donne le plan principal du domaine construit. On peut distinguer en première approximation deux nuages de points.

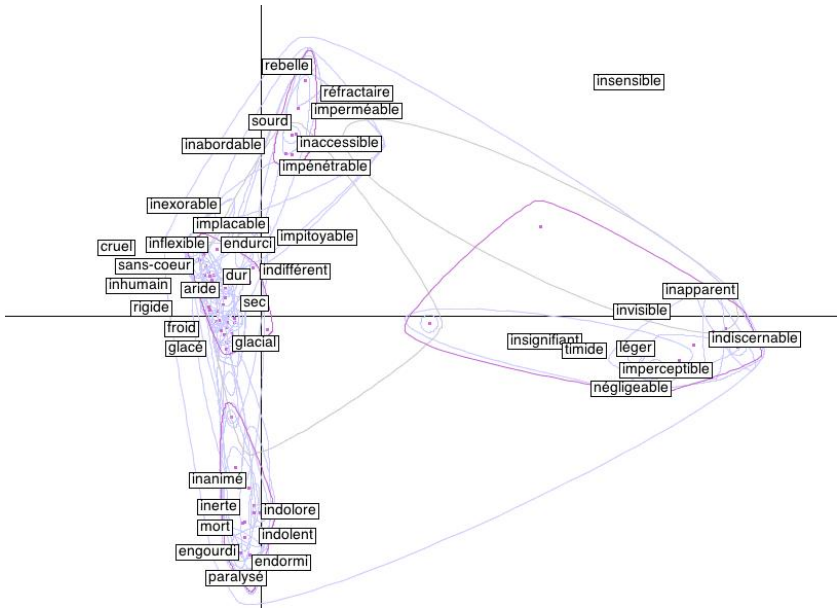


Figure 7: Projection sur le plan principal de domaine associé au mot *insensible*.

Le premier, vertical, contient les valeurs morale (au centre) et perceptive qui s'appliquent le plus souvent à une personne. Un second nuage, plus nettement séparé, contient la valeur perceptive et qui s'applique à un phénomène externe.

En somme, l'espace géométrique permet aussi de construire une organisation du sens d'un item lexical. Le domaine associé à un mot est centré, quand elle existe, autour d'une valeur générique, les valeurs proches du point de vue sémantique sont contiguës et enfin des valeurs sémantiques homonymiques ou quasi-homonymiques sont nettement distinguées. Enfin, le modèle a la capacité de construire le domaine associé, non plus à une seule entrée, mais à une liste d'entrées et de rendre compte des phénomènes de recouvrement du sens entre les différents mots de la liste.

# 3 Un modèle de composition sémantique développé dans le cadre formel des espaces topologiques des AS

Notre objectif est d'utiliser la représentation géométrique du lexique pour calculer la sémantique des combinaisons de mots afin (i) de sélectionner pour chaque mot les valeurs sémantiques pertinentes et donc de déterminer la valeur du mot dans un contexte donné ; (ii) de positionner dans un même espace l'expression composée et des expressions connues du système et qui lui sont

proches du point de vue de l'interprétation.

Comme Pustejovsky [Pustejovsky, 1998], nous avons choisi de travailler sur les syntagmes nominaux et verbaux. Les effets de sens produits par la variation des mots en contexte sont présentés, dans cet article, à travers les syntagmes nominaux contenant l'adjectif *rapide* et les syntagmes verbaux contenant le verbe *couper*. Ces deux mots sont souvent cités dans la littérature pour leur complexité polysémique. Voici quelques exemples introductifs de paraphrases associées à des expressions les contenant (certains de ces exemples sont tirés du *Le Trésor de la langue française informatisé* (TLFI), [Dendien, 2002]) :

- Un fleuve rapide (qui coule à fort débit) / Une route rapide (dans laquelle les virages ont été aménagés et qui permet aux véhicules de circuler sans ralentir) / Un cheval rapide (qui se déplace à une vitesse élevée) / Un esprit rapide (qui assimile, qui réagit promptement) / Une médication rapide (qui produit son effet dans un court délai).
- Couper un arbre (abattre un arbre, scier un arbre) / Couper une communication (fermer une communication) / Couper un abcès (ouvrir un abcès).

Afin de construire un modèle de composition, nous avons besoin d'une représentation des items lexicaux (dont les principes sont rappelés en section 2), d'une représentation de la structure des expressions et enfin de la définition de l'opérateur de composition lui-même. Nous présentons donc maintenant les données syntagmatiques utilisées. À la suite de quoi nous définirons une distance entre syntagmes puis l'opérateur de composition.

## 3.1 La base des syntagmes

Nous avons constitué un corpus de syntagmes construits à partir d'expressions courantes qui font état des emplois les plus usuels et des constructions les plus fréquentes qu'un locuteur peut produire ou entendre. Ces expressions ont été construites à partir d'un corpus contenant des textes journalistiques et littéraires et sélectionnées en fonction des emplois donnés dans des dictionnaires comme le Robert [Robert, 1996] ou le TLFI [Dendien, 2002]. Nous donnons ici des syntagmes associés au verbe *couper : couper le pain, couper du pain, couper le cordon, couper les ongles, couper la viande, couper bras et jambes, couper un chat, couper les cheveux à quelqu'un, couper une communication téléphonique, couper l'appétit, couper le souffle, couper l'eau, etc.* et au terme *rapide : un rythme rapide, une décision rapide, une pellicule rapide, une voie rapide, etc.* Le tableau 1 fait état pour un ensemble de dix items lexicaux du nombre de ses synonymes dans la base lexicale et du nombre de syntagmes de la base d'expressions qui le contiennent.

## 3.2 Une distance entre syntagmes

Dans ce travail nous proposons de considérer que la granularité sémantique des cliques s'apparente à celles des expressions. En effet, le contenu d'une clique

| Entrée lexicale | Nombre de synonymes et parasynonymes de la base lexicale | Nombre de syntagmes de la base de syntagmes contenant l'entrée lexicale |
|---|---|---|
| *couper* | 148 | 97 |
| *échancrer* | 8 | 2 |
| *moissonner* | 13 | 10 |
| *retrancher* | 54 | 8 |
| *supprimer* | 80 | 18 |
| *traverser* | 31 | 23 |
| *actif* | 48 | 15 |
| *rapide* | 60 | 23 |
| *subit* | 12 | 4 |
| *vif* | 131 | 18 |
| Moyenne sur l'ensemble des données constituées | 34 | 10 |

Table 1: Nombre de synonymes et de syntagmes pour un échantillon d'entrées lexicales.

associée à un verbe représente une valeur qui contraint le choix du complément possible. Ainsi, par exemple, pour le verbe *couper* :

- la clique "*amputer*, *censurer*, *couper*, *retirer*, *retrancher*, *supprimer*" renvoie à des compléments de type production (*censurer un ouvrage*, *retrancher un paragraphe*, etc.)
- la clique "*couper*, *scier*, *tronçonner*" renvoie à des compléments constitués d'un objet solide (*scier un tronc d'arbre*, etc.)
- la clique "*couper*, *mouiller*, *tremper*" renvoie à des compléments de type boisson (*mouiller du vin*, etc.)

De la même façon, un complément ou un qualificatif contraint la sémantique de la tête du syntagme nominal. Pour l'adjectif *rapide* :

- la clique "*abrupt*, *brusque*, *escarpé*, *raide*, *rapide*" qualifie une voie, chemin, etc. dont la topologie entraîne pour celui qui l'emprunte un déplacement rapide (*un sentier escarpé*, *une descente rapide*, etc.).
- la clique "*brusque*, *fulgurant*, *rapide*, *violent*" qualifie un processus et éventuellement par extension une entité pouvant développer ce type de processus (*un parcours fulgurant*, *un virus fulgurant*, etc.).
- La clique "*bref*, *compendieux*, *concis*, *court*, *rapide*, *sommaire*, *succinct*" qualifie le plus souvent un ouvrage écrit ou oral (*un discours sommaire*, etc.).

Cette analogie entre granularité des cliques et contraintes sur les compléments conduit à choisir comme représentation d'un syntagme un point dans un espace multidimensionnel. Les coordonnées des syntagmes et leurs distances respectives sont calculées en tenant compte des différents facteurs suivants : leur structure syntaxique, les listes des lemmes et des mots de fonction qu'ils contiennent. Le tableau 2 représente ces données pour quelques expressions associées au mot *couper*.

| Syntagme | Structure syntaxique | Mots pleins (lemmes) | Mots de fonction |
|---|---|---|---|
| *couper le pain* | $ST_1$ | *couper, pain* | *le* |
| *couper le pain en tranches* | $ST_4$ | *couper, pain, tranche* | *le, en* |
| *couper les communications téléphoniques* | $ST_2$ | *couper, communication, téléphonique* | *les* |
| *couper les branches de l'arbre* | $ST_3$ | *couper, branche, arbre* | *les, de, l'* |

| | |
|---|---|
| $ST_1$ | [V[Det, N]] |
| $ST_2$ | [V[Det, [N, Adj]]] |
| $ST_3$ | [V[Det,[N,[Prep,[Det,N]] |
| $ST_4$ | [V[Det,N][Prep,N]] |
| etc. | |

Table 2: Description des données utilisées pour la constitution des espaces syntagmatiques.

Les coordonnées d'un ensemble de syntagmes sont calculées par application d'une analyse factorielle des correspondances effectuée sur la matrice comportant en lignes les expressions et en colonnes les structures syntaxiques, les cliques associées aux mots pleins et aux mots de fonction. Comme pour la construction des domaines lexicaux, chaque élément de la matrice contient un 0 ou un 1 suivant l'appartenance du paramètre situé sur une colonne au syntagme situé sur une ligne de la matrice. La distance entre expressions est calculée à partir des coordonnées de cet espace. Elle permet, par la suite, de définir le voisinage d'une expression et de repérer une expression donnée par rapport l'ensemble des expressions déjà connues (voir figures 9 et 10).

# 4 Un opérateur de composition guidé par la topologie des espaces sémantiques

L'opérateur de composition a pour but (i) de sélectionner les régions pertinentes du domaine sémantique associé à de chaque constituant d'un syntagme donné ; (ii) de calculer les proximités sémantiques entre ce syntagme et les autres syntagmes en les plaçant dans un espace sémantique commun. L'opérateur

utilise comme entrées les domaines sémantiques associés aux mots constituant le syntagme-requête ainsi que l'espace des syntagmes qui contiennent ces mots et leurs synonymes. Il fournit en sortie les sous-domaines sémantiques pertinents associés aux constituants et le sous-espace des syntagmes les plus proches du point de vue du sens. Cet opérateur est défini par la topologie des espaces suivant un principe de continuité ici énoncé :

> Soit la requête constituée de la liste des mots : $Mot_1, Mot_2$, soient $E_{Mot_1}$ et $E_{Mot_2}$ les domaines sémantiques construits suivant le modèle présenté dans le paragraphe 2.1.3 autour des mots $Mot_1$ et $Mot_2$, soit $E_{Expr}$ l'espace des syntagmes contenant les termes de $E_{Mot_1}$ et de $E_{Mot_2}$, soit $P$ l'application qui a un terme d'un domaine sémantique lexical ($E_{Mot_1}$ ou $E_{Mot_2}$) associe l'ensemble des syntagmes de $E_{Expr}$ qui contiennent cet élément, alors un voisinage $V_1 \subseteq E_{Mot_1}$ de rayon $\epsilon_1$ et un voisinage $V_2 \subseteq E_{Mot_2}$ de rayon $\epsilon_2$ et un voisinage $W \subseteq E_{expr}$ de rayon $\epsilon$ sont pertinents pour la composition sémantique si $P(V_1) \bigcap P(V_2) \subseteq W$ [4] (voir figure 8).

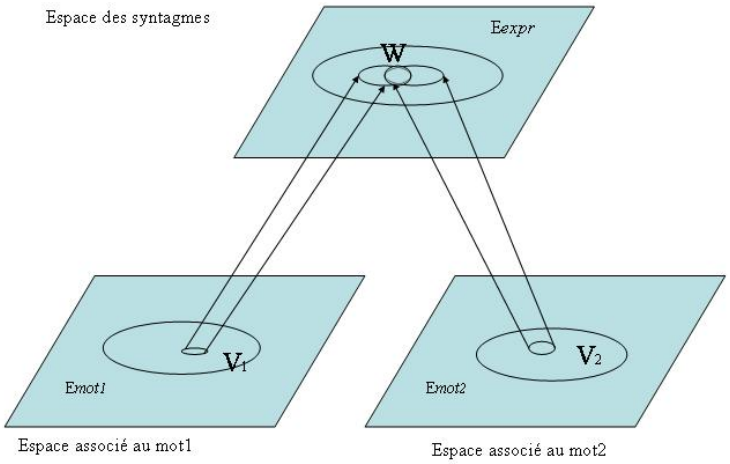


Figure 8: Schéma de sélection sémantique entre espaces lexicaux et syntagmatique

[4] $\epsilon$ , $\epsilon_1$ et $\epsilon_2$ sont des paramètres du modèle qui mesurent la diffusion du processus. Des valeurs faibles renvoient à des réponses précises ; le système, en augmentant ces paramètres (et donc en élargissant les voisinages), donne accès à la fois à des termes et à des expressions plus distantes.

# 5 L'algorithme

Le système prend en entrée une expression. Il en extrait la liste des mots pleins et lui associe, comme valeur initiale, sa structure syntaxique (par exemple [V [Det,N]] pour un syntagme verbal, [Det,N,Adj] pour un syntagme nominal. Voici la suite des étapes ainsi que les résultats obtenus pour la requête *couper l'herbe*.

- Étape 1 : Construction pour chaque mot plein composant la requête de son espace sémantique associé. Pour l'exemple, deux espaces sont construits, celui qui contient le domaine associé à *couper* et celui qui contient le domaine associé au mot *herbe*, (figure 9).
- Étape 2 : Construction de l'espace des syntagmes contenant des termes similaires à tous les mots pleins de la requête. Cet espace comporte entre autres des expressions comme *fermer l'entrée du champ* ou *étendre le champ de ses expériences* car *fermer* et *étendre* appartiennent à l'espace associé à *couper* et *champ* à celui de *champ*, (figure 9).
- Étape 3 : Application de l'opérateur de composition.
- Étape 4 : Reconstruction de la géométrie des espaces lexicaux et de l'espace des syntagmes à partir des éléments sélectionnés. On remarquera ici que les termes appartenant au domaine initial de *couper* comme *diluer, censurer, blesser...* ne sont pas retenus. Seuls restent des termes pertinents pour l'expression *couper de l'herbe* comme *arracher, faucher...*, figure 10.
- Étape 5 : Affichage des expressions pertinentes. Dans l'exemple, deux types de phrases proches de la requête initiale figurent dans ce nouvel espace, celles qui expriment un arrachage comme *ôter les mauvaises herbes* et celles qui renvoient à l'idée de faucher un pré ou encore de tondre le gazon (l'herbe étant alors comprise comme une étendue), (figure 10).

# 6 Étude des résultats pour les syntagmes verbaux

Nous avons fait varier dans les requêtes les compléments associés au verbe *couper*. Nous reproduisons dans les tableaux suivants les résultats obtenus. La première colonne comprend la requête, la seconde et la troisième des synonymes ou parasynonymes de chacun des mots constituants la requête et qui sont sélectionnés par le calcul. La dernière colonne comporte une liste de syntagmes proposés par le système et qui ont participé à la sélection et qui sont donc proches de la requête.

## 6.1 Sélection des valeurs

Le tableau 3 montre que suivant ses compléments, le verbe *couper* désigne l'arrêt d'un processus (comme dans le cas du complément *communication*), la séparation d'une unité (comme pour le complément *arbuste*) ou encore la traversée (pour le complément *chemin*) et que ces valeurs sont correctement sélectionnées par le modèle.

| Requête | Synonymes et parasynonymes du premier constituant sélectionnés | Synonymes et parasynonymes du second constituant sélectionnés | Syntagmes sélectionnés |
|---|---|---|---|
| couper un arbuste | couper, tailler, amputer, blesser, mutiler, raccourcir, retrancher, rogner, tronquer, ébrancher, écharper, élaguer, émonder, | arbre, arbuste | couper l'arbre<br>ébrancher un arbre<br>mutiler un arbre<br>tailler un arbre<br>tronquer un arbre |
| couper une communication | couper, arrêter, interrompre, rompre, barrer, suspendre, fermer, intercepter, cesser, terminer, finir, briser | communication, transport, circulation, correspondance, rapport, relation, fréquentation, liaison, message, dépêche | couper les communications,<br>couper une communication téléphonique,<br>fermer les communications,<br>interrompre la circulation,<br>interrompre les communications<br>briser une correspondance,<br>intercepter un message, |
| couper une conversation | arrêter, cesser, couper, fermer, finir, interrompre, terminer, trancher | conversation, débat | interrompre une conversation téléphonique, terminer un débat |
| couper un chemin | couper, traverser | chaussée, chemin, | couper le chemin à quelqu'un,<br>traverser la chaussée |

Table 3: Exemples de résultats obtenus par variation des compléments du verbe *couper*.

## 6.2 Interprétation du sens

Il n'existe pas toujours d'expression contenant la liste des termes de la requête et les expressions sélectionnées peuvent même ne contenir aucun de ces termes. Pour la requête *couper un arbuste* (tableau 3), le système ne contient pas d'expression pertinente contenant le terme *arbuste* qui, dans ce cas, est compris, du point de vue du processus *couper* comme un analogue d'*arbre*. En ce sens, la sélection des termes relève d'un processus d'interprétation. Ce processus peut aussi servir à interpréter des requêtes composées de termes qui semblent mal associées. Nous donnons ici, (tableau 4), un exemple de résultat

obtenu à partir d'une association mal assortie : *couper l'air*. Cette expression n'existe pas en français. Cependant le modèle propose une interprétation qui ramène l'expression mal construite à des expressions idiomatiques connues comme *déchirer l'air* ou *fendre la bise* qui décrivent un mouvement très rapide [5].

| | | | |
|---|---|---|---|
| couper l'air | couper, déchirer, fendre, rompre | air, bise, souffle, vent | déchirer l'air,<br>fendre l'air,<br>fendre la bise, |

Table 4: Un exemple d'interprétation pour une association non usuelle.

### 6.3 Voisinages lexicaux et traits sémantiques

On pourrait envisager de montrer que les voisinages sémantiques matérialisent la réalisation ou la sélection d'un trait sémantique. Cependant nous voulons ici mettre en évidence que le système a la capacité de choisir les bons voisinages même si les compléments partagent plusieurs traits et que la composition sémantique ne se fonde pas sur les traits communs. Voici mis en comparaison (tableau 5) les résultats pour les requêtes *couper l'eau* et *couper le vin*. Bien que l'eau et le vin soient des liquides et des boissons, dans le premier cas l'*eau* est conçue comme une denrée amenée dans une maison par un réseau (comme le gaz ou l'électricité) et le vin comme une boisson susceptible d'être diluée.

| | | | |
|---|---|---|---|
| couper l'eau | barrer, boucher, couper, fermer, obstruer | eau | couper l'eau,<br>fermer l'eau,<br>boucher une voie d'eau,<br>boucher une conduite d'eau |
| couper le vin | couper, mouiller, mélanger, mêler | vin | couper son vin,<br>mélanger des vins,<br>mêler le vin,<br>mouiller son vin |

Table 5: Traits sémantiques et composition : un exemple.

## 7 Étude des résultats pour les syntagmes nominaux

Ici nous avons fait varier les termes qualifiés par l'adjectif *rapide*. Comme précédemment, les colonnes 2 à 4 des tableaux 6 et 7 représentent les résul-

[5] Notons que les résultats du modèle pourraient être exploités comme aide à la rédaction ou à la traduction pour un locuteur étranger ne maîtrisant pas les expressions idiomatiques de la langue.

tats du calcul.

#### 7.0.1 Sélection des valeurs

Le terme *rapide* convoque principalement deux propriétés : la brièveté et l'intensité. Les exemples du tableau 6 montrent comment se fait la sélection en fonction de la tête du syntagme. De plus, l'emploi de l'adjectif a un effet sur la signification des termes qu'il qualifie. Il sélectionne de façon préférentielle le caractère fonctionnel, procédural et temporel. Ainsi *un chemin rapide* est un itinéraire rapide.

| Requête | Synonymes et parasynonymes du premier constituant sélectionnés | Synonymes et parasynonymes du second constituant sélectionnés | Syntagmes sélectionnés |
|---|---|---|---|
| un chemin rapide | chemin, voie, avenue, sentier, artère, chaussée, couloir, route, trajet, direction, itinéraire, ligne, course méthode, moyen, manière | rapide, large, actif | une course rapide,<br>une route large,<br>une voie rapide,<br>une méthode active |
| un procédé rapide | pratique, procédure, procédé, façon, manière, marche, moyen, méthode, tactique, technique, art, attitude, comportement, conduite, allure, style | rapide, prompt, vif, brusque, ardent, actif, violent, turbulent, soudain, pressant, court, bref, concis,<br>pressé, expéditif, diligent, | une méthode active,<br>une vive allure<br>un procédé sommaire<br>une procédure expéditive<br>d'une manière vive,<br>d'une manière soudaine,<br>d'une façon pressante,<br>d'une manière pressante,<br>d'une manière brève<br>un style vif |
| un récit rapide | narration, récit | rapide, sommaire | une narration sommaire |

Table 6: Résultats obtenus par variation des noms auxquels se rapporte l'adjectif *rapide*.

### 7.1 Interprétation du sens

Nous donnons ici (tableau 7), comme précédemment, le résultat obtenu pour une requête ne figurant pas parmi les expressions usuelles (*un cerveau rapide*) et dont le résultat figure la capacité du modèle à créer une interprétation dans un contexte inédit. Ici, *cerveau* est compris comme la fonction qui lui est associée : *intelligence, jugement*, etc. et *rapide* désigne l'aptitude à fonctionner promptement *brillant, vif*, etc.

| | | | |
|---|---|---|---|
| un cerveau rapide | cerveau, entendement, esprit, génie, intelligence, jugement, jugeote, raison, tête | actif, agile, alerte, ardent, expéditif, fringant, hâtif, leste, pressé, preste, prompt, rapide, sémillant, vif | un esprit alerte<br>un esprit rapide et brillant<br>un jugement expéditif<br>un jugement hâtif<br>une intelligence vive |

Table 7: Exemple d'interprétation d'une association comportant le terme *rapide*.

# 8 Évaluation des résultats obtenus.

Afin de tester la pertinence des résultats obtenus, nous avons constitué pour chacun des deux types de syntagmes (V-N), (N-ADJ) une liste de mots sur lesquels portent soit le verbe *couper* soit l'adjectif *rapide*. Ces deux listes ont été construites à partir des cooccurrences les plus fréquentes données par le logiciel d'interrogation de la base Frantext sur un corpus de textes du XIXème et du XXème siècles. De l'ensemble des cooccurrences du mot *couper* et du mot *rapide*, nous n'avons retenu que les termes qui sont susceptibles d'apparaître dans un syntagme de type V-N (resp. N-ADJ ou ADJ-N.). Ainsi, par exemple, les mots *laisser* et *travers* de la liste cooccurrences du verbe *couper* sont retirés, les mots *tête* et *liens* sont conservés. De même pour l'adjectif *rapide*, les mots *grand* et *maintenant* sont retirés, les mots *développement* ou *départ* conservés. Après cette première sélection, en avons fait une seconde pour retenir finalement les termes qui ont une fréquence de cooccurrence supérieure à 10 avec *couper* (resp. à 40 avec *rapide*). Les deux listes, notées $L_{couper}$ et $L_{rapide}$, contiennent 57 mots.

Le modèle a ensuite été lancé sur l'ensemble des requêtes de type *"coupe(r) (Det) $t_i$(s)"* (où Det$\in \{le, la, les, l', un, des, []\}$, $t_i \in L_{couper}$, i<=57) et sur l'ensemble des requêtes de type *"(Det)$t_j$(s) rapide(s)"* ($t_j \in L_{rapide}, j <= 57$). Ces requêtes correspondent à la colonne 1 des tableaux de résultats (3 à 7). Les valeurs de $\epsilon$ qui définissent le voisinage d'application de l'opérateur de composition ont été fixées de façon globale pour chacune des deux listes de requêtes. Ces paramètres étant fixés et comme cela a été précédemment décrit, le modèle donne comme résultats pour chaque requête trois ensembles de réponses : deux ensembles de synonymes ou parasynonymes pertinents pour la requête (un ensemble par mot plein de la requête) et un ensemble de syntagmes permettant l'application de l'opérateur de composition. Afin d'évaluer la pertinence de ces réponses, nous avons testé leur emploi en interrogeant le Web grâce au moteur de recherche Google de la manière suivante. Pour chaque requête initiale, un synonyme ou parasynonyme $Syn_{mot}$ d'un des mots pleins de la requête est noté 1 si à une requête de type

$$"coupe(r)\ (Det)\ Syn_{t_i}(s)"ou"Syn_{couper}\ (Det)\ t_i(s)"$$

$$(resp."(Det)\ Syn_{t_j}(s)\ rapide(s)"ou"(Det)\ t_j(s)\ Syn_{rapide}(s)")$$

le moteur Google retourne, dans la liste de ses 100 premières réponses, des syn-

tagmes qui sont des paraphrases de la requête initiale *"coupe(r) (Det) $t_i$(s)"* (resp. *"(Det) $t_i$(s) rapide(s)"*). Par exemple, pour le syntagme initial *"une action rapide"* , le synonyme *brusque* de *rapide* appartenant à la liste des réponses du modèle a été noté 1 car le moteur de recherche renvoie l'emploi attesté *"une action brusque"* sémantiquement proche de la requête initiale. De même, le synonyme *mouvement* de *action* a été noté 1 car le moteur de recherche permet d'attester l'emploi *"un mouvement rapide"* très similaire à cette même requête initiale. En revanche, le mot *fringant*, synonyme de *rapide* et donné par le modèle comme réponse pour le même syntagme *"une action rapide"* reçoit la note 0 car il n'existe pas d'emploi attesté par Google de syntagme du type *"(Det) action(s) fringantes(s)"*. De plus, s'il existe les emplois attestés mais qui ne sont proches sémantiquement de la phrase initiale, le synonyme sélectionné par le modèle reçoit la note 0. Enfin, l'ensemble des syntagmes donnés par le modèle (quatrième colonne dans la présentation des résultats des tableaux 3 à 7) sont également notés (1 ou 0) suivant leur proximité sémantique au syntagme initial (première colonne). Le tableau 8 récapitule cette évaluation.

| | Substantifs sur lesquels porte *rapide* | *rapide* | Syntagmes |
|---|---|---|---|
| Nombre total (nombre moyen par requête) d'items sélectionnés par le modèle | 506 (8,9) | 522 (9,2) | 298 (5,2) |
| Nombre total (proportion) d'emplois attestés et pertinents | 418 (83%) | 431 (83%) | 223 (75%) |
| | *couper* | Substantif complément de *couper* | Syntagmes |
| Nombre total (nombre moyen par requête) d'items sélectionnés par le modèle | 720 (12,6) | 1386 (24,3) | 1014 (17,8) |
| Nombre total (proportion) d'emplois attestés et pertinents | 595 (83%) | 998 (72%) | 703 (69%) |

Table 8: Résultats de l'évaluation du modèle à partir de syntagmes de type *"(Det) N rapide(s)"* et de syntagmes de type *"coupe(r) (Det) N(s)"*. Les valeurs des paramètres de l'opérateur de composition sont globalement fixés : $\epsilon_1 = 0,06, \epsilon_2 = 0,05, \epsilon = 0,05$ pour les syntagmes nominaux et $\epsilon_1 = 0,05, \epsilon_2 = 0,05, \epsilon = 0,05$ pour les syntagmes verbaux.

# 9 Discussion et perspectives

Nous avons montré comment, à partir d'une base initialement non structurée de synonymes, d'une base réduite de syntagmes prototypiques et par usage du paradigme géométrique, le modèle permet de rendre compte de la variabilité et de la créativité du sens en contexte en positionnant un syntagme-requête par rapport aux syntagmes les plus semblables de la base et en sélectionnant les valeurs sémantiques des unités lexicales qu'il contient. De plus, ce modèle offre différents avantages par rapport aux modèles cités dans la première partie de l'article. La distinction et l'organisation des valeurs sémantiques des unités lexicales sont automatiques (ce qui n'est pas réalisé dans WordNet) ; elles donnent d'aussi bons résultats pour les noms, les verbes que les adjectifs. Les marqueurs qui permettent d'interpréter le sens d'une expression sont automatiquement proposés par le modèle : ceux sont les synonymes pertinents pour l'opération de composition. Ceci est un avantage par rapport au modèle de Kintsch dans lequel ces termes sont choisis à la main par l'auteur. Enfin, le modèle ne nécessite pas la définition *a priori* d'une liste de traits ou de primitives comme c'est le cas dans le LG. Pour deux types de phrases nous avons évalué les résultats obtenus par le modèle en les comparant aux emplois attestés disponibles sur le Web et renvoyés par le moteur de recherche Google. Cette évaluation effectuée à partir du verbe *couper* et de l'adjectif *rapide* donne dans chacun des cas des résultats satisfaisants alors que les paramètres du modèle sont fixés de façon globale. En effet, comme dans toute application continue la taille des voisinages pertinents (et donc la valeur d'$\epsilon$) dépend localement de la topologie de l'espace sur laquelle elle porte. Une évaluation plus précise nécessiterait de calculer pour chaque requête la taille optimale des voisinages qui permettra de majorer le nombre de bonnes réponses tout en minorant le nombre de réponses erronées. Pour cela, il serait nécessaire de tenir compte de la structure individuelle des domaines sémantiques en interaction et plus précisément des densités locales des points (une zone de forte densité représente une valeur sémantique assez homogène alors qu'une diminution de la densité représente souvent le passage d'une valeur sémantique à une autre; la valeur des paramètres $\epsilon$ devra être calculée pour s'adapter à cette densité locale en empêchant l'incorporation de valeurs sémantiques non pertinentes). Cette optimisation effectuée permettra d'évaluer la capacité du modèle à fournir le plus grand nombre de bonnes réponses tout en rejetant le plus grand nombre de mauvaises[6].

Enfin un autre élément demanderait à être pris en compte : celui de l'indépendance relative entre la qualité de la réponse du système et la note obtenue par interrogation d'un corpus. En effet, si on peut vraisemblablement adhérer à l'idée qu'un emploi attesté dans un corpus est un bon indicateur de la pertinence de la réponse, en revanche, il n'est pas exclu qu'un emploi non attesté ne soit pas valide. Cette dernière remarque rejoint le problème précédemment

[6]Notons toutefois que sur la base d'un sous-échantillon de 10 requêtes initiales prises au hasard, la proportion de synonymes non pertinents et non sélectionnés par rapport à l'ensemble des synonymes non pertinents est d'environ 90%. Pour les syntagmes, cette proportion est de l'ordre de 70%.

cité de la créativité du sens : certains emplois pourraient apparaître et ainsi valider *a posteriori* des réponses du système. Dans le paragraphe qui suit, nous traçons quelques perspectives qui découlent de l'observation des résultats de cette première étude.

Dans plusieurs études [Rouibah et al., 2001, Ji, 2004], nous avons testé avec succès la pertinence cognitive du paradigme géométrique. Cependant, dans le cas de la sémantique des phrases et afin de s'approcher au plus près d'un fonctionnement véritablement cognitif, il serait cohérent que l'acquisition de la base lexicale et de la base d'expressions prototypiques dont dépend le modèle soit elle aussi proche de celle des humains. La base des expressions, est de petite taille, contrairement aux corpus qui servent à LSA. Elle est constituée d'exemples les plus fréquemment utilisés dans la langue et ressemble donc aux phrases entendues par un individu moyen. En revanche la base lexicale a été constituée à partir de l'expertise de lexicographes (voir [Ploux, 1997] pour le mode de constitution de cette base). Afin d'obtenir une base à la fois plus exhaustive et plus évolutive, nous envisageons d'utiliser non pas des jugements d'experts, mais directement des corpus de textes en retenant pour synonymes des termes qui partagent des contextes d'emplois similaires à la manière dont cela a été déjà réalisé par Grefenstette [Grefenstette, 1994]. Nous pensons qu'une base ainsi construite permettrait une plus grande précision des résultats. Nous chercherons également à compléter cette première étude par un travail sur des expressions syntaxiques plus complexes que celles ici envisagées.

Au delà du travail théorique portant sur la recherche de modèles appropriés à la langue nous nous attacherons à promouvoir les résultats en les intégrant à la plateforme des AS afin d'offrir des fonctionnalités étendues. Sur la base des remarques formulées par les utilisateurs, nous savons d'ores et déjà que les résultats de ce travail répondent à un besoin : nombre d'entre eux nous ont demandé d'étendre le modèle lexical à un modèle des expressions. En plus de son apport en modélisation, le présent travail est donc aussi une réponse à cette demande. Une perspective à plus long terme serait, comme cela a été fait au niveau lexical, de proposer un modèle d'appariement sémantique entre mots et expressions de deux langues différentes. L'étude ici présentée en constitue la première étape.

# References

[Benzécri, 1980] Benzécri, J.-P. (1980). *L'analyse des données : l'analyse des correspondances*. Bordas, Paris.

[Burgess and Lund, 1997] Burgess, C. and Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, 12:177–210.

[Chuquet and Paillard, 1987] Chuquet, H. and Paillard, M. (1987). *Approche linguistique des problèmes de traduction anglais-français*. Ophrys, Paris.

[Dendien, 2002] Dendien, J., editor (2002). *Le Trésor de la langue française informatisé*. ATILF-CNRS Université Nancy 2, http://atilf.atilf.fr/tlfv3.htm.

[Dowty et al., 1981] Dowty, D., Wall, R., and Peters, S. (1981). *Introduction to Montague Semantics*. D. Reidel Publishing Company, Dordrecht.

[Geach and Black, 1960] Geach, P. and Black, M., editors (1952, 1960). *Translations from the Philosophical Writings of Gottlob Frege*. Blackwell, Oxford.

[Godart-Wendling et al., 1998] Godart-Wendling, B., Ildefonse, F., Pariente, J.-C., and Rosier-Catach, I. (1998). Penser le principe de compositionnalité : éléments de réflexion historiques et épistémologiques. *Traitement automatique des langues*, 39(1):9–34.

[Grefenstette, 1994] Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer.

[Ikeda, 1998] Ikeda, S. (1998). Manual response set in a stroop-like task involving categorization of english and japanese words indicates a common semantic representation. *Percept Mot Skills*, 87(2):467–474.

[Illes and Francis, 1999] Illes, J. and Francis, W. (1999). Convergent cortical representation of semantic processing in bilinguals. *Brain and Language*, 70(3):347–363.

[Jackendoff, 1999] Jackendoff, R. (1999). *Semantics and Cognition*. MIT Press, Cambridge, Massachusetts.

[Ji, 2004] Ji, H. (2004). *A Computational Model for Word Sense Representation Using Contextual Relations*. Mémoire de thèse en sciences cognitives.

[Kintsch, 2001] Kintsch, W. (2001). Predication. *Cognitive Science*, 25:173–202.

[Landauer et al., 1998] Landauer, T. K., Foltz, P., and Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

[Ploux, 1997] Ploux, S. (1997). Modélisation et traitement informatique de la synonymie. *Linguisticae Investigationes*, 21(1):1–28.

[Ploux and Ji, 2003] Ploux, S. and Ji, H. (2003). A model for matching semantic maps between languages (french/english, english/french). *Computational Linguistics*, 29(2):155–178.

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

[Pustejovsky, 1998] Pustejovsky, J. (1998). *The Generative Lexicon*. MIT Press, Cambridge, Massachusetts.

[Robert, 1996] Robert, L., editor (1996). *Le Petit Robert. Dictionnaire de la langue française.* Havasinteractive.

[Rouibah et al., 2001] Rouibah, A., Ploux, S., and Ji, H. (2001). Un modèle spatial des représentations lexicales impliquées dans la reconnaissance des mots écrits. In H. Paugam-Moisy, V. Nyckees, J. C. P., editor, *La cognition entre individu et société.* Hermès.

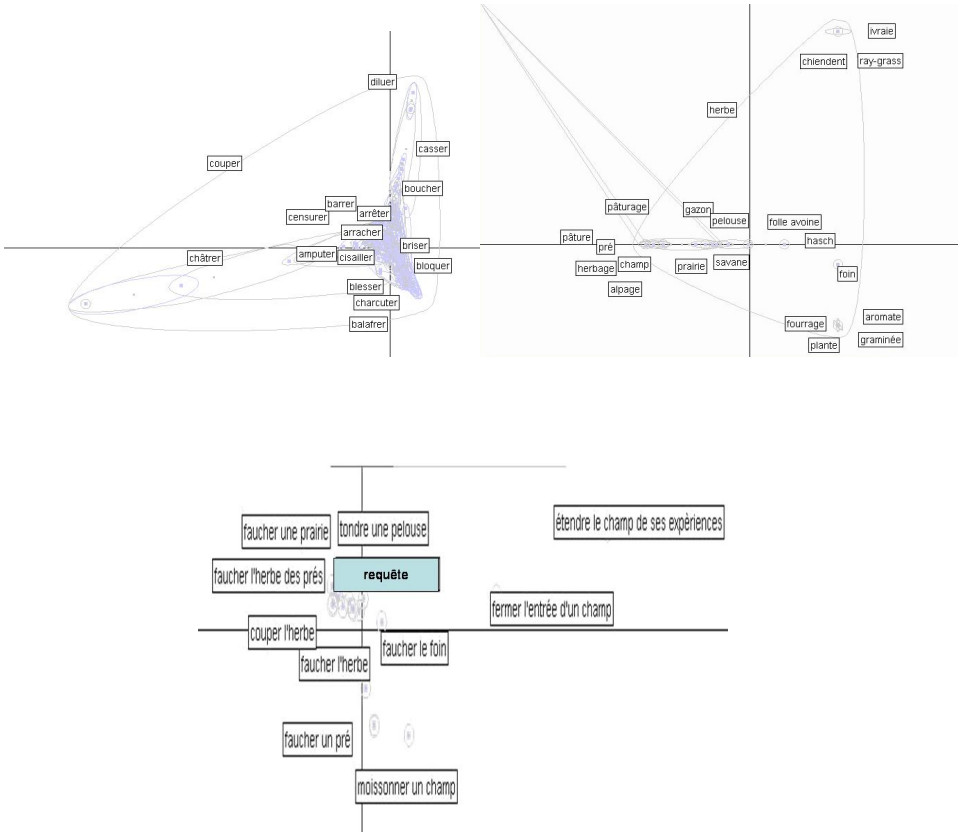


Figure 9: Les différents espaces construits pour les termes *couper* et *herbe* et pour les syntagmes associés à ces deux termes avant application de l'opérateur de composition (étape 1 et étape 2 de l'algorithme). On pourra noter que dans cet exemple, l'expression requête *couper l'herbe* appartient déjà à la base d'exemple.

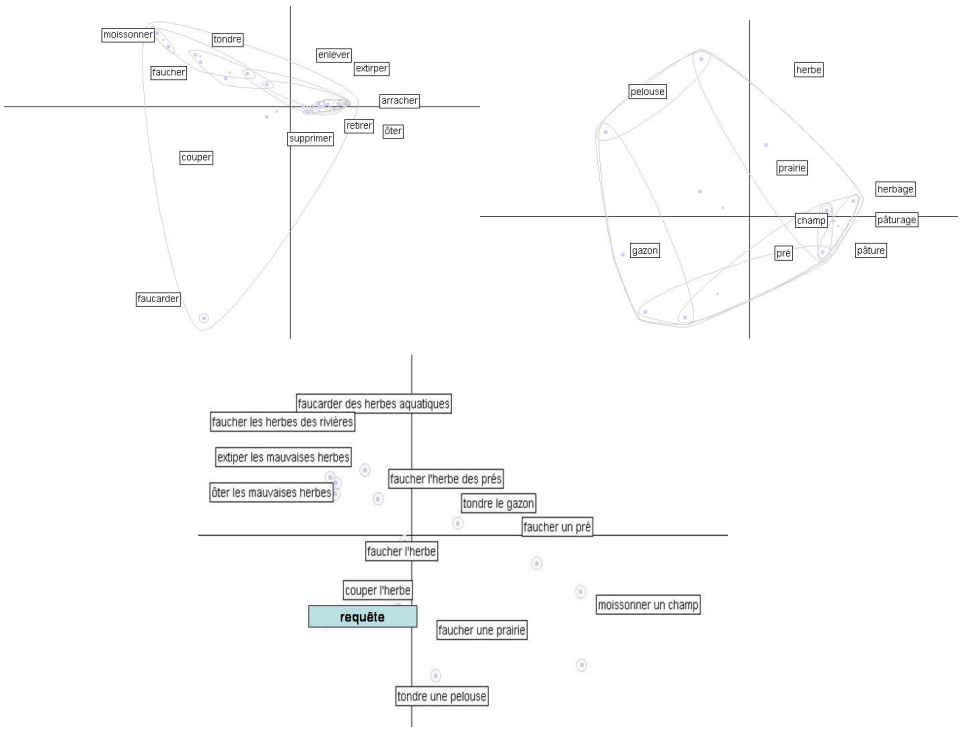


Figure 10: Les différents espaces construits après la sélection effectuée grâce à l'opérateur de composition (étape 4) respectivement pour le terme *couper*, pour le terme *herbe* et pour la requête *couper l'herbe*