Modèles et sémantique lexicale

Sabine Ploux

Modèles et sémantique lexicale

Sabine Ploux

1 Introduction

Comprendre une phrase, interpréter un texte, traduire, communiquer... sont autant d'activités cognitives qui supposent que les mots ont un sens véhiculé à travers leurs emplois. La question du sens a souvent pu paraître, par nature, moins objectivable que d'autres domaines linguistiques comme celui de la syntaxe. Certains auteurs pensent même qu'elle résiste à toute tentative de modélisation computationnelle. Cette réticence à la formalisation tient à une appréciation d'inadéquation entre d'une part l'objet de la sémantique et d'autre part les cadres mathématiques couramment utilisés et les implémentations dont ils ont fait l'objet. Nous commencerons par introduire très rapidement une partie de la sémantique : celle qui a donné lieu à des travaux de modélisation. Nous préciserons ensuite ce que nous entendons par modélisation. Ceci nous permettra d'aborder la question d'une éventuelle adéquation. Pour cela, nous nous limiterons à l'étude des modèles en sémantique lexicale à travers l'examen de quelques-uns d'entre eux choisis pour représenter la diversité des cadres formels dans lesquels ils ont été développés. Les résultats de la modélisation seront ensuite envisagés de deux points de vue : celui du traitement automatique des langues et celui de leur pertinence cognitive.

2 Sémantique, modélisation et informatique

Sémantique La sémantique est un champ d'étude polymorphe. En linguistique, la sémantique recouvre de façon non exhaustive :

- l'étude du sens lexical. Les méthodes mises en œuvre pour cela utilisent le plus souvent :
  - les proximités lexicales : synonymie (*logement* est un synonyme d'*habitation*), hypo et hyperonymie (*chat* est un hyponyme d'*animal* et inversement *animal* est un hyperonyme de *chat*), métonymie (dans l'expression *boire un verre*, le contenant *verre* est métonyme de la boisson qu'il contient)...
  - la décomposition des mots en traits sémantiques (/*transport*/ est un trait sémantique partagé par les mots *métro, train, autobus,* etc.. /*ferré*/ est un trait qui différencie *autobus* de *train* (exemple tiré de [Rastier, 1987]) ;
- l'étude du sens des combinaisons de mots, des phrases ou des textes. Ce domaine comprend le plus souvent :
  - le calcul de la signification d'une combinaison de mots en contexte (*un bon livre* est un livre qui se lit avec plaisir, *un bon marcheur* est une personne qui parcourt de longues distances avec facilité; le sens d'un adjectif peut donc varier en fonction du nom sur lequel il porte);
  - le calcul de la signification de la phrase ou du discours par l'étude de leur forme logique. Ce calcul comprend en particulier la détermination des conditions de vérité d'un énoncé mais aussi le calcul de la portée sémantique des unités de la phrase à l'œuvre dans les anaphores, la négation, la quantification...

En psycholinguistique, ce domaine recouvre généralement l'étude de la représentation des concepts en mémoire. Les concepts lexicaux étudiés appartiennent le plus souvent à des catégories larges comme la distinction vivant (animaux, plantes, parties du corps...)/non vivant (outils...)
La diversité entre les différents types d'approche témoigne d'un état à la fois créatif et transitoire de la recherche. La plupart des auteurs ([Gharbia et al., 1998,
, Moeschler and Auchlin, 2000]) s'accordent cependant sur le fait que la sémantique est l'étude des significations et que « la signification est quelque chose qui n'est pas du langage » ([Moeschler, 2004]). On a donc une mise en relation de deux espaces de nature hétérogène : celui des signes linguistiques d'une part et celui d'un autre domaine prélinguistique d'autre part. Cet autre domaine est variable. Il peut être celui des notions, des concepts ou des référents (individus, choses, événements ou états du monde), des conditions de vérité... Pour certains auteurs, ce domaine substrat est cognitif, pour d'autres il renvoie directement aux objets du monde. Se pose alors la question de l'existence d'un modèle formel qui permettrait de synthétiser et de théoriser ce lien.
Modélisation Un exemple paradigmatique de la recherche de modèles est la naissance du calcul différentiel au XVIIème siècle. Ce cadre a permis de répondre aux attentes de la physique et plus précisément à celui de la théorisation du mouvement des corps. Cet exemple montre que pour aborder un champ disciplinaire, il peut être nécessaire de créer des idéalités mathématiques. Actuellement, en sciences du langage, il n'existe pas de consensus sur le choix du cadre formel à adopter. On peut citer le recours à des modèles logico-algébriques comme l'ont initialement proposé Chomsky en syntaxe [Chomsky, 1969] ou Montague en sémantique [Montague, 1974, Dowty et al., 1981] ou encore des modèles qui utilisent la théorie des graphes, les espaces vectoriels, la théorie des singularités, la géométrie, les systèmes dynamiques... Ce manque de convergence résulte peut être d'une part de l'absence de modèle tout à fait adéquat et qui resterait à construire et d'autre part du fait que peu d'auteurs se saisissent de la question si ce n'est à travers une dispute entre les tenants des modèles discrets et ceux des modèles continus.
Informatique L'informatique est la troisième composante des travaux en sémantique. Elle a permis, quand cela a été possible, l'implémentation des modèles, donné lieu au développement d'outils, élargi le champ des expérimentations linguistiques. On pourra se reporter à [Habert et al., 1997] pour un panorama très complet. Comme tout médiateur, l'informatique représente une contrainte. En particulier, nous essaierons de montrer que la relation entre le modèle et son implémentation informatique est une conversion qui impose des limites. Ces limites sont plus ou moins fortes suivant l'adéquation entre cadre formel et machine numérique. Cependant, si l'outil n'est jamais le prolongement fidèle d'une intention théorique, il ouvre l'accès à de nouvelles possibilités : la gestion d'immenses masses de données textuelles a permis l'émergence d'une linguistique expérimentale aussi appelée linguistique de corpus. Enfin, la recherche de modèles computationnels en sémantique est un champ d'étude très productif aujourd'hui. Car de ces modèles dépend la performance des systèmes d'acquisition et de représentation des connaissances, de recherche d'information, de traduction automatique...

3 Quelques grands courants de la modélisation en sémantique lexicale

Les principaux cadres formels utilisés en sémantique lexicale sont (1) les graphes (2) la logique mathématique (et en particulier le lambda-calcul) (3) les espaces vectoriels (4) les réseaux neuronaux (5) les espaces topologiques (6) les systèmes dynamiques. Certains de ces

cadres sont plus spécifiquement adaptés à la représentation du sens (il s'agit des graphes, des espaces vectoriels et des espaces topologiques), d'autres s'attachent aussi à rendre compte du processus de calcul ou de constitution du sens (il s'agit de la logique mathématique, des réseaux de neurones, des systèmes dynamiques). Et c'est dans cet ordre que nous les présenterons.

4 Graphes et représentation des connaissances lexicales

Plusieurs systèmes (WordNet [Fellbaum, 1998b], FrameNet [Ruppenhofer et al., 2005], ..) utilisent le graphe comme paradigme de représentation lexicale. L'exemple le plus connu reste certainement le thésaurus WordNet. Un des objectifs fondateurs de WordNet était de proposer un thesaurus exhaustif des connaissances lexicales dont toute personne dispose. Le paradigme revendiqué par les auteurs pour bâtir cette base lexicale prend appui sur une approche dite classique (ou aristotélicienne) caractérisée par la représentation d'un concept par une liste de propriétés nécessaires et suffisantes. Par exemple, le concept de *rouge-gorge* se caractérise par les propriétés : *a la poitrine rouge, est un vertébré, à sang chaud, a un bec, des ailes, des plumes, peut chanter, voler, pondre des œufs*. Cette représentation par propriétés induit une hiérarchie fondée sur l'héritage : (*rouge-gorge* hérite des propriétés de tous ses concepts hyperonymes : *oiseau, animal...*). La figure 1 reprend la présentation que Miller [Miller, 1998] fait de cette hiérarchie conceptuelle.
Le formalisme des graphes a été utilisé pour représenter cette organisation. C'est un formalisme discret à valeurs finies fondé sur deux notions : l'unité conceptuelle et la relation entre unités couplées respectivement à celle de nœud et d'arête d'un graphe. La hiérarchie fondatrice du système basée sur l'hyperonymie a ensuite été augmentée d'autres relations comme la méronymie qui lie une partie de l'objet au tout ou encore l'antonymie qui lie des opposés, la troponymie : un verbe V_1 est un troponyme de V_2 si V_1 est une façon particulière de faire V_2. Par exemple, *cheminer* est une troponyme de *marcher* [Fellbaum, 1998a]. Notons aussi que dans WordNet le concept lexical n'est pas représenté par un mot mais par une liste de synonymes (synset). Ceci permet de séparer plusieurs concepts désignés par un même nom. Ainsi le mot *house* est décomposé en 12 concepts lexicaux dont *firm, house, business firm* ou encore *family, household, house, home, menage*.
Du point de vue de son implémentation informatique, Wordnet est un projet d'envergure : il contient actuellement 117597 synsets. Cependant l'architecture hiérarchique reste plus adaptée à la représentation des noms qu'à celles des verbes, et *a fortiori* des adjectifs et des adverbes. Enfin, la constitution de la base lexicale WordNet est peu automatisée, elle a été largement effectuée « à la main ».

5 Proximité sémantique et modèles vectoriels

Ces dernières années, plusieurs modèles vectoriels ont vu le jour comme LSA [Landauer et al., 1998] ou HAL [Burgess and Lund, 1997]. Le développement des modèles vectoriels fait écho à une mouvance alternative à l'approche classique des concepts. Cette alternative prend appui sur le constat d'une inadéquation entre hiérarchie conceptuelle et résultats obtenus lors d'expériences avec des sujets. Ainsi, [Smith et al., 1974] ont mis en évidence des différences de temps de traitement entre les énoncés du type : *A robin is a bird, (Un moineau est un oiseau)* et des énoncés du type : *A penguin is a bird, (un pingouin est un oiseau)*. Le premier type est traité plus rapidement que le second. Ces résultats impliquent que l'appartenance à une même catégorie, ici *bird*, n'est pas calculée sur la base d'une unique liste de conditions nécessaires et suffisantes (ou dans la représentation hiérarchique par la distance entre le

concept hyponyme et le concept hyperonyme). En effet, la liste des conditions étant identique, le temps de traitement devrait être le même. Et de même, le nombre d'arcs séparant *robin* et *penguin* de *bird* étant le même dans WordNet, le temps de traitement devrait rester inchangé. Afin d'expliquer ces phénomènes, il a été proposé que les concepts ne sont pas organisés en mémoire à la manière d'une hiérarchie conceptuelle mais plutôt autour de prototypes [Rosch, 1983] et suivant leur similarité sémantique. *Moineau* ou *rouge-gorge* sont des prototypes d'*oiseau*, en revanche, sur un gradient de prototypie, *pingouin* est relativement éloigné de ces éléments centraux.
En adéquation avec cette mouvance alternative en pyscholinguistique et contrairement à WordNet, les modèles vectoriels ne présupposent pas une organisation *a priori* des concepts ou de la sémantique des unités lexicales. Les auteurs mettent même en doute la possibilité de parvenir à définir le sens des mots (voir [Kintsch, 2001]). Dans cette perspective, les liens sémantiques sont établis à partir des emplois relevés dans des corpus de textes. Ces corpus peuvent contenir plusieurs millions de mots.
Les expériences de psycholinguistique utilisent comme paramètre d'investigation les temps de réaction ou des jugements de similarité fournis par des sujets. Il semble naturel que ces indices qui prennent des valeurs numériques variables trouvent dans la notion de distance dans un espace vectoriel un support adapté. Afin de calculer les distances sémantiques, LSA établit une matrice comportant en colonnes les paragraphes (ou les phrases) du corpus traité et en ligne les mots. Chaque case $m_{i,j}$ de cette matrice contient initialement la valeur 0 ou 1 suivant que le mot_i appartient ou non au $paragraphe_j$. Une analyse factorielle[1] est ensuite effectuée sur cette matrice initiale. Le résultat est une nouvelle matrice qui associe à chaque mot ses coordonnées dans un espace multidimensionnel. La proximité entre deux vecteurs associés à deux mots est définie mathématiquement par leur cosinus : $\cos(vecteur_{mot1}, vecteur_{mot2})$. Cette méthode permet de faire la synthèse des liens de cooccurences mais aussi d'utiliser le fait que deux mots qui ont des contextes similaires sont aussi sémantiquement proches. Si dans le corpus le mot *roman* apparaît plusieurs fois dans le voisinage du mot *livre*, et qu'il en est de même du mot *poésie*, aussi alors LSA permettra de repérer à la fois une proximité sémantique entre les mots *livre* et *roman* ou *livre* et *poésie* mais aussi entre les mots *roman* et *poésie*.
Différentes caractéristiques séparent cette approche de WordNet :

- Les liens sémantiques entre unités lexicales dépendent du corpus utilisé et ne sont donc pas fixés *a priori*. Le mot *connaissance* n'aura pas les mêmes voisins sémantiques selon qu'on aura choisi un corpus général ou un corpus de spécialité en philosophie.
- Les structures initiales du corpus sont : le mot, la phrase, le paragraphe et le texte, tous repérés par des séparateurs : blanc, ponctuation, saut de ligne... La donnée de ces unités, même très basiques, constitue cependant un choix crucial puisque, dans cette perspective, *pomme de terre*, par exemple, n'est pas constitué d'une mais de trois unités lexicales, et que *maison* et *maisons* sont deux unités lexicales distinctes.
- Enfin, la représentation associée à un mot est atomique (comme l'est aussi un synset), mais aussi unique en ce sens que les différentes composantes du vecteur ne sont pas directement interprétables en caractéristiques sémantiques séparées : dans ce formalisme, le vecteur constitue une unité indécomposable. Il en découle, que les valeurs sémantiques d'un mot ne sont pas représentées pour elles-mêmes, seule est donnée une liste de mots voisins qui, bien qu'associés chacun à une ou plusieurs de ces valeurs ne permettent pas de les distinguer. Par exemple, les plus proches voisins de *party* construits à partir du corpus *General_Reading_up_to_12th_Grade* sur le site http://lsa.colorado.edu/ sont par ordre de proximités décroissantes : (le chiffre

correspond à la valeur du cosinus) 0,83 parties, 0,73 prohibitionists, 0,73 prohibitionist, 0,73 spokesperson, 0,72 democrats, 0,71 antifederalists, 0,67 sorauf, 0,67 chairpersons, 0,66 tuba, 0,66 democratic, 0,66 partisanship, 0,65 nominating, 0,65 candidates, 0,65 invite, 0,65 birthday, 0,64 railwaymen, 0,64 eec, 0,63 whig, 0,63 factions. Les voisins relatifs à *party* au sens de parti politique sont mêlés à ceux relatifs à *party* au sens de fête.

Enfin, les modèles vectoriels sont tout à fait adaptés à une implémentation informatique. Leur mise en oeuvre ne nécessite qu'un corpus de taille suffisante et des algorithmes matriciels relativement classiques.

6 Voisinages et classification sémantique, l'alternative géométrique

Il existe peu de modèle géométrique. On pourra trouver chez P. Gärdenfors [Gärdenfors, 2000] un argumentaire philosophique en leur faveur. Gärdenfors après avoir énoncé les limites à la fois des approches symboliques et des approches connexionnistes, propose le paradigme géométrique comme un pont entre ces deux niveaux de description. Dans cette perspective, l'espace associé à un concept est construit à partir d'un ensemble de dimensions perceptives. Par exemple, Fairbanks et Grubb 1961 [Fairbanks and Grubb, 1961] donnent une représentation géométrique des voyelles construites comme des aires délimitées par un polygone sur un plan dont les axes sont définis par les valeurs des deux premiers formants[2]. Deux éléments caractérisent les modèles géométriques : la séparation de l'espace des formes de celui du contenu, la projection d'une forme sur un domaine de l'espace du contenu cognitif. Dans le domaine lexical, le choix d'un modèle géométrique permet à la fois de représenter une structure interne des concepts par interprétation des différentes zones qui constituent le domaine et de rendre compte de la similarité sémantique entre unités. Cette similarité est définie par la mesure du recouvrement entre des aires qui leur sont associées et aussi par la distance qui les sépare. Ces deux caractéristiques font la synthèse d'une capacité à classifier (WordNet propose une classification *a priori* des différents sens des mots), et à mesurer les proximités comme dans les modèles vectoriels. De plus, le paradigme géométrique est un support de simulation de notre capacité à traduire le sens. En effet, les tentatives pour construire un modèle translinguistique se sont révélées difficiles dans le paradigme hiérarchique (WordNet) et ont été peu développées dans le paradigme vectoriel. En revanche, dans le paradigme géométrique, la définition d'un espace de contenu permet de modéliser les phénomènes de traduction lexicale. La figure 3 donne une idée de ce que peuvent être les différentes projections associées à deux langues, voir aussi [Ploux and Ji, 2003]. Cependant, un problème posé par les modèles géométriques réside dans le fait qu'il est difficile d'avoir accès directement au contenu sémantique. Dans le modèle des *Atlas sémantiques* ([Ploux, 1997, Ploux and Victorri, 1998]) cet inconvénient a été déjoué par un artefact utile : la notion de clique[3]. En effet, si la représentation associée à la sémantique d'un mot est, comme nous l'avons supposé, un domaine dans un espace multidimensionnel, il est nécessaire pour construire ce domaine, de représenter les unités qui le composent. Les cliques ont initialement été calculées à partir de la relation de synonymie (on peut aussi proposer des cliques de contexte [Ji et al., 2003]). Les cliques de synonymie sont des ensembles maximaux de mots tous synonymes les uns des autres. Dans ce modèle, une clique est représentée par l'intersection des différents domaines associés à la liste des synonymes (voir la figure 4). Une

implication de la propriété de maximalité est qu'il n'existe aucun autre mot dans la langue qui puisse diviser cette intersection. Pour cette raison, une clique représente une unité minimale de sens, un « grain » de sens. Voici énoncées quelques propriétés de ces unités :

- Les cliques contrairement aux synsets ne sont pas des unités de langage ni donc d'un métalangage. Il est difficile à la fois de les désigner et de nommer leurs différences. On pourra s'en persuader à travers l'exemple des trois cliques suivantes contenant le mot *insensible* :
    - cruel, dur, féroce, impitoyable, implacable, inexorable, inhumain, insensible
    - cruel, dur, impitoyable, implacable, inexorable, inflexible, inhumain, insensible
    - cruel, dur, impitoyable, implacable, inexorable, inflexible, insensible, sévère
- Il existe une topologie sous-jacente à l'ensemble des cliques associées à un mot qui permet de distinguer des valeurs et de passer par des chemins continus d'une valeur à une autre qui lui est proche. Ainsi, les cliques données ci-dessus sont des exemples de cliques relatives à la valeur « morale » du mot *insensible* ; les cliques
    - endormi, engourdi, inerte, insensible
    - engourdi, froid, inerte, insensible

  sont associées à la valeur « physique » ; les cliques

    - imperméable, impénétrable, inaccessible, insensible, réfractaire, sourd
    - imperméable, impénétrable, inaccessible, insensible, réfractaire, étranger

  à une valeur qu'on pourrait qualifier d'« émotionnelle » ; les cliques

    - imperceptible, inapparent, insensible, invisible
    - imperceptible, indiscernable, insaisissable, insensible, invisible

  à une valeur « perceptuelle » qui contraire aux précédentes valeurs ne désignent pas une personne mais un phénomène externe. L'examen de l'ensemble des cliques met en évidence l'existence de chemins de cliques dans lesquels une clique partage au moins un mot avec la suivante, et qui font passer de façon progressive d'une valeur à une autre.

La construction de la forme associée au mot initial, ici *insensible*, permet de faire la synthèse de l'ensemble des liens de proximité. Pour cela, de façon similaire à ce que fait LSA sur une matrice de paragraphes et de mots, on utilise une analyse factorielle des correspondances [Benzécri, 1980] sur la matrice qui comprend des cliques en ligne et des mots en colonne. Cette méthode permet de calculer les coordonnées des cliques représentées par des points dans un espace multidimensionnel. Les mots eux sont représentés par l'enveloppe des points-cliques qu'ils contiennent. Enfin, un algorithme de classification permet de distinguer à partir du nuage de points formé par les cliques les différentes valeurs du mot. La figure 5 donne le résultat pour le mot *insensible*.
Ce résultat met en évidence la capacité du modèle (i) à déterminer une valeur générique (quand elle existe) ; cette valeur est positionnée près de l'origine des axes (ii) à déterminer des valeurs proches et des valeurs homonymiques ou quasi-homonymiques qui sont nettement séparées des autres sur la carte. En somme, la modélisation géométrique est une modélisation continue qui associe à un mot non plus un atome ou plusieurs atomes de sens (vecteur ou

noeud d'un graphe) mais un domaine qui permet la représentation de l'organisation de ses différentes valeurs sémantiques.
Comme les modèles précédents, les modèles géométriques sont des modèles de représentation qui, sans outils supplémentaires, ne rendent pas compte des processus de calcul du sens ou de la forme argumentale ou schématique d'une unité lexicale.

7 Grammaire du sens et modèles logico-algébriques

Il existe une tradition de modélisation logico-algébrique en linguistique. Afin de théoriser l'aptitude humaine à produire des phrases grammaticales, Chomsky a décrit la syntaxe comme un système formé d'un vocabulaire, d'axiomes et de règles d'inférence. Montague a intégré ce formalisme syntaxique à la modélisation de la sémantique des phrases. Plus récemment J. Pustejovsky [Pustejovsky, 1998] a utilisé un cadre du même type pour l'étude de la sémantique lexicale.
Le *Lexique génératif* (LG) de J. Pustejovsky est une proposition-cadre motivée par l'inadéquation d'une conception énumérative du sens comme l'est WordNet. Les trois arguments majeurs de la critique des théories à valeurs finies sont leur incapacité à rendre compte :

- de la possible créativité du sens d'un mot dans un contexte inédit (Pustejovsky prend pour exemple l'adaptation du sens des adjectifs comme *good* (*bon*), voir exemple plus haut° ;
- du partage possible du sens des mots (par exemple des verbes *bake* (*cuire*), *cook* (*cuisiner*) ou *fry* (*frire*)) ;
- des multiples réalisations syntaxiques des mots (par exemple, le verbe *forget* (*oublier*) pour lequel les différents types de compléments déterminent l'interprétation sémantique : *oublier d'où l'on vient*, c'est oublier la réponse à la question sous-jacente, contrairement à *oublier son parapluie* qui ne met pas en jeu une question).

L'idée est de remplacer la donnée d'un ensemble de valeurs sémantiques fixées *a priori* par une capacité calculatoire à déterminer le sens en contexte. Pour réaliser ce projet J. Pustejovsky a choisi le lambda-calcul. Ce choix prolonge l'entreprise de la grammaire générative par la détermination des phrases non plus seulement syntaxiquement mais aussi sémantiquement bien formées. Comme en syntaxe, la détermination du sens en contexte est réalisée par un ensemble d'axiomes et des règles de dérivation. Les axiomes ici sont l'ensemble des unités lexicales munies d'une structure de type attribut-valeur. Cette structure comprend (i) un composant argumental (par exemple le verbe *bake* a une structure argumentale formée de deux éléments : le premier est de type animé, le second est de type « massif ») , (ii) des composants événementiels (trois éventualités : l'état, le procès et la transition), (ii) des composants de type *qualia*, et (iv) des liens d'héritage au sein du réseau lexical. Les composants de type *qualia* comportent eux-mêmes quatre aspects :

- un aspect constitutif : la relation entre l'objet et ses composantes ;
- un aspect formel qui distingue l'objet d'un domaine plus large ;
- un aspect télique qui décrit la fonction de l'objet (*eat* (*manger*) pour le mot *cake* (*gâteau*)) ;
- un aspect agentif qui donne les facteurs impliqués dans la création de l'objet (l'acte de cuire (désigné par *bake-act* dans le LG) pour *cake*).

Trois opérations permettent de calculer l'interprétation des mots en contexte et d'assurer le caractère bien formé des combinaisons. Ces opérations rendent compte, par exemple, des différences entre la phrase *bake the potatoes* dont le résultat correspond à un changement d'état des pommes de terre alors que la phrase *bake a cake* signifie la création de l'objet gâteau. Cette proposition est séduisante pour gérer le traitement de la polysémie. Cependant, le choix d'un cadre logique imposerait pour être tout à fait adapté que soient assurées (1) la cohérence du système, qui requiert qu'on ne puisse pas aboutir à une contradiction par application des opérations et (2) sa complétude, ce qui signifie que tous les énoncés puissent être dérivés des axiomes (ici l'ensemble des représentations des lexèmes) par application des opérations. De plus, d'un point de vue méthodologique, et donc aussi informatique, la mise en oeuvre de ce paradigme nécessite aussi la spécification des valeurs des composants de la structure pour tous les éléments du lexique (seuls quelques exemples typiques sont traités dans [Pustejovsky, 1998]]), et sur la détermination de l'ordre d'application des règles.

8 Processus de converge du sens et réseaux de neurones

Les modèles connexionnistes sont nés d'une analogie biologique. On ne cherche plus à modéliser, comme dans les modèles symboliques, les objets de la pensée mais son substrat (les neurones et leurs connexions en réseau), afin de simuler l'activité cognitive.
Différents réseaux de neurones ont été proposés pour modéliser les traitements liés au lexique. Certains ont pour objet la modélisation du fonctionnement linguistique comme le réseau de B. Victorri et C. Fuchs [Victorri et Fuchs, 1996] qui cherche à rendre compte de la polysémie adverbiale en contexte, d'autres tentent de simuler des processus mentaux comme le réseau de M. Masson [Masson, 1995] qui modélise l'amorçage sémantique. Parmi les arguments en faveur du choix d'un modèle neuronal figure celui de la nécessité d'une représentation distribuée des objets (ici les mots ou les concepts). Dans l'approche distribuée, un concept est représenté par un ensemble d'unités correspondant à différents types de processus. Cette approche s'oppose à l'approche par réseaux sémantiques, qualifiés d'approches locales, dans laquelle chaque concept est représenté par un nœud d'un graphe, objet insécable et sans parties, relié aux autres concepts par les arêtes du graphe. Dans la représentation proposée par M. Masson, une partie des composantes du vecteur associé à chaque mot encode les traits phonologiques, une autre les traits orthographiques, la troisième les traits sémantiques. À chaque pattern d'entrée est associé un vecteur souvent composé de 0 et de 1 (1 pour la présence du trait, 0 sinon). Cependant, une démarche assez courante, comme chez M. Masson, consiste à allouer une certaine portion du vecteur à la représentation d'un ensemble de traits et à les remplir de façon aléatoire pour un pattern donné ce qui positionne ces travaux plus du côté d'une simulation virtuelle que d'une modélisation réelle.
Les modèles distribués permettent une certaine tolérance : la donnée partielle en entrée des caractéristiques du concept (comme sa forme orthographique) donne accès, par convergence du réseau, au concept désigné. Un concept (ou le sens d'un mot) sera donc finalement représenté dans le réseau par le centre d'un bassin d'attraction sur l'espace constitué des différentes dimensions. Un bassin d'attraction est, comme le montre la figure 6, un puits de potentiel séparé par des lignes de crêtes. Les différents bassins d'attractions sont tous séparés et le réseau modélise donc un processus de choix ou de décision entre différentes valeurs possibles (nommer un objet dans [Masson, 1995], déterminer la valeur sémantique en contexte dans [Victorri et Fuchs, 1996]). Or, en sémantique, on observe des phénomènes de recouvrement entre unités lexicales (voir les arguments de J. Pustejovsky donné au paragraphe précédent). La séparation en bassins, sans intersection mutuelle, ne permet pas de représenter ces recouvrements (la représentation interne du sens lexical restant en effet cachée). De plus, dans les implémentations le fait que le nombre de patterns (3 paires de mots

dans [Masson, 1995]) reste souvent dérisoire au regard de ceux couramment manipulés par un être humain, contrairement aux modèles de représentation (vectoriels, hiérarchiques ou géométriques), constitue une limite actuelle de ces modèles.

9 Morphogenèse du sens et théorie des singularités

R. Thom [Thom, 1977, Thom, 1980] a défendu l'idée que le langage comme la morphogenèse du vivant peut se comprendre comme une relation entre un substrat doté de propriétés géométriques et dynamiques et des morphologies observables qui sont les catégories linguistiques. Cette approche diffère de l'approche précédemment qualifiée de géométrique en ce sens que la relation entre ces deux strates n'est pas une relation d'association mais le résultat d'un processus décrit en fonction de paramètres de contrôle et qui détermine le nombre et les caractéristiques des morphologies de surface.
Dans le cadre de cette théorie, dénommée théorie des catastrophes, R. Thom a proposé une typologie de la sémantique des verbes par utilisation de la théorie des singularités pour la classification des différentes interactions possibles entre arguments [Thom, 1980]. La figure 7 reprend l'exemple des verbes *couper*, *séparer* et *traverser*. Cette approche était novatrice à plusieurs titres. Elle modélise la nature dynamique du contenu sémantique des verbes indépendamment d'un métalangage, dont R. Thom rejette l'usage car il aboutirait à des problèmes de circularité. Elle fait découler le sens de la régulation, de la stabilisation de conflits agissant sur supports continus, supports ultimes du signifié.
Cependant, il s'agit plus d'un modèle associé à un concept lexical qu'à un mot. En effet, les actants en interaction sont assimilés à des points ; ceci ne permet pas de comprendre les différences sémantiques qui se produisent quand on opère une substitution. Par exemple, le verbe *couper* ne décrit pas la même interaction dans *couper du pain* ou dans *couper la route* (dans le premier cas, le pain est divisé, dans le second, la route reste inchangée et le schéma le plus adapté est donc celui associé au verbe *traverser*). La nature propre des arguments est donc un paramètre nécessaire à l'accès au contenu sémantique. Si la théorie des singularités peut rendre compte de la déformation qui fait passer d'une configuration à une autre, comme par exemple ici de la configuration de *couper* à celle de *traverser*, il n'en reste pas moins que le mode de déformation en fonction de la nature des mots en composition n'est pas décrit.
De plus, le nombre de verbes d'une langue étant bien supérieur au nombre d'interactions listées dans le cadre de la théorie des singularités, il semble nécessaire de complexifier, mais il n'est pas dit comment, ces schémas pour rendre compte de l'immense diversité lexicale. Ce paradigme a inspiré d'autres travaux comme celui de J. Petitot [Petitot-Cocorda, 1985] qui l'a étendu à la morphogenèse du sens pour rendre compte de la double articulation du langage et des structures narratives. Cependant, peu de réalisations informatiques en sont issues. Le manque de propositions inspirées de la théorie des systèmes dynamiques tient probablement à la difficulté pour une machine actuelle d'être une machine à singularités. En effet, il faut, pour appliquer les résultats mathématiques, traduire des valeurs numériques en formules analytiques. Le passage du numérique à l'analytique ainsi que le calcul des singularités qui en découle constituent des obstacles majeurs à l'utilisation des ordinateurs pour mettre en œuvre ce type de modèles.

10 Modélisation, sémantique et traitement automatique

Les enjeux Internet a permis une immense accessibilité aux données textuelles provenant du monde entier et disponibles dans un nombre croissant de langues. Pour pouvoir les utiliser, il faut des systèmes d'interrogation performants. L'espoir, dans les années qui viennent, est de

rendre possible une recherche d'information qui aille au-delà d'une simple interrogation par mots et qui traite directement le contenu sémantique de la requête. À ce projet s'ajoute celui d'une recherche d'information multilingue [Grefenstette, 1998] qui permettra à un utilisateur d'exprimer sa requête dans sa langue maternelle et d'avoir accès aux documents disponibles dans une autre langue. Tout comme la traduction automatique ou l'aide à la traduction automatique, la recherche d'information multilingue suppose une capacité à déterminer un contenu sémantique commun entre deux langues. Du point de vue lexical, les problèmes soulevés par ces différents domaines sont la levée des ambiguïtés lexicales, la gestion de la polysémie, du contexte, mais aussi le traitement morphologique des mots complexes.

Modèles sémantiques et informatique : quel est le mariage réussi ? Les réalisations actuelles qui se révèlent les plus pertinentes pour répondre à ces enjeux sont les réseaux lexicaux : thésaurus, ontologies (voir [Vossen, 2003] par exemple). Les modèles vectoriels ont également été utilisés. Les réseaux neuronaux ou les modèles génératifs, faute de bases lexicales suffisantes sont moins présents. La tendance est donc plus à l'extraction de relations lexicales, à l'acquisition et à la constitution de grandes bases qu'à la proposition de modèles qui dériveraient beaucoup à partir de peu. Le nombre d'articles portant sur ces questions et utilisant des méthodes statistiques d'analyse des corpus en témoignent. L'informatique, grâce aux immenses capacités de mémoire, s'est révélée plus utile pour stocker, chercher et repérer des formes ou encore calculer des proximités que pour dériver, inférer, etc. Les modèles qui « collent » aux caractéristiques des machines ont donc été privilégiés. Dans ce contexte, les efforts de modélisation proprement linguistique n'ont pas été véritablement récompensés. Le gain que les systèmes informatiques apportent réside plus dans leur capacité utile à extraire des corpus d'exemples qui alimenteront la démarche théorique que dans leur capacité à la mettre en œuvre.

11 Modélisation, sémantique et cognition

Différents types d'expériences permettent d'interroger la représentation de la sémantique des mots en mémoire. À ces expériences s'ajoute l'étude des déficits. Nous présentons ici certains modes d'analyse utilisés en montrant comment les résultats sont comparés aux modèles précédemment présentés.

Une organisation par catégories Plusieurs études font état, chez certains patients aphasiques, de déficits affectant sélectivement une catégorie qu'elle soit large comme l'ensemble des entités vivantes vs. non vivantes ou encore plus fines comme les parties du corps, ou les fruits et végétaux [Forde and Humphreys, 2002]. Le patient est alors incapable d'accéder aux informations relatives aux concepts de la catégorie déficitaire. Le choix d'une organisation hiérarchique des concepts comme celle proposée par WordNet permettrait d'expliquer ces pathologies. En effet, la rupture d'un lien dans l'arborescence des concepts rend inaccessibles tous les sous-concepts comme cela semble se produire chez ces patients.

Une organisation par proximités L'étude de la représentation du sens des mots en mémoire a fait l'objet de théories et plus récemment d'expériences. Comme nous l'avons rappelé plus haut, aux catégories définies par des propriétés nécessaires et suffisantes se sont opposées des propositions prenant pour support le prototype ou encore le gradient d'appartenance. Plusieurs études ont comparé résultats expérimentaux et résultats computationnels. [Lund et al., 1996] ont comparé avec succès amorçage sémantique et associatif au modèle vectoriel HAL. L'amorçage sémantique consiste à évaluer le lien entre un mot amorce présenté dans un laps de temps très court au sujet et un mot cible. L'hypothèse est que si le traitement du mot amorce a un effet différentiel sur le traitement du mot cible alors on peut en déduire qu'en mémoire ces deux traitements sont liés. Pour cela, on étudie le temps de réaction à une tâche demandée au sujet qui peut être par exemple une tâche de décision lexicale (le sujet doit dire

si le mot cible est un mot ou un non mot). [Laham, 1997] a également montré que dans LSA les couples de mots de type nom-catégorie (comme *pomme-fruit*) ont des distances sémantiques significativement plus petites que des couples non appariés (*pomme-animal*). [Vigliocco et al., 2004] ont montré que la mesure de similarité sémantique calculée par le modèle vectoriel FUSS permet de rendre compte des résultats obtenus dans une expérience au cours de laquelle le sujet devait nommer rapidement une image (dans ce cas le mot erroné produit est aussi un voisin sémantique donné par le modèle) ou encore dans des expériences d'amorçage sémantique. Enfin, [Ji et al., 2008] ont montré que le modèle ACOM, modèle géométrique construit à partir de relations de contexte, produit des voisins sémantiques comparables aux mots produits par des sujets dans une expérience d'association de mots.

Une organisation distribuée mettant en jeu un substrat sensori-moteur Enfin, un tournant s'est opéré dans la recherche sur les aires cérébrales impliquées dans le traitement du sens des mots : initialement localisé par la communauté scientifique dans des régions spécifiques situées autour de la scissure sylvienne de l'hémisphère gauche, ce traitement apparaît, au regard de nouveaux résultats, distribué suivant des réseaux corticaux dont la topographie reflète la sémantique du mot traité. Ainsi [Haul et al., 2004] ont montré que le traitement des mots d'action impliquant la face, le bras ou la jambe (comme les verbes *to lick, to pick, to kick* (*lécher, cueillir, donner un coup de pied*)) dans une tâche de lecture passive activent des aires adjacentes ou se superposant aux aires impliquées respectivement dans le mouvement de la langue, de la main ou du pied. Ces résultats montrent l'implication d'aires associées à des traitements de plus bas niveau et indépendants d'un système linguistique et appuient l'idée d'un substrat sémantique sensori-moteur. Ces expériences privilégieraient donc la plausibilité des modèles géométriques ou dynamiques pour ce qui est du lien avec un contenu infra-linguistique. Enfin, la distribution de ces réseaux fait écho à une modélisation neuronale.

12 Quelles perspectives ?

En somme, les différents modèles s'adaptent à un aspect du traitement ou de la représentation sémantique des mots mais rencontrent des difficultés à intégrer l'ensemble des aspects. L'atomisme associé à des modèles formels (graphes ou espace vectoriels) permet des réalisations à large couverture lexicale (WordNet ou LSA) mais ne permet pas de rendre compte d'une organisation logique du sens. Cette logique du sens est le coeur des modèles génératifs, mais l'appariement entre le processus et le modèle choisi pose des problèmes de validation. Les réseaux de neurones, s'ils répondent à la distribution révélée par l'imagerie cérébrale, ne rendent pas compte des phénomènes de recouvrements sémantiques. Enfin, les modèles hiérarchiques, vectoriels et géométriques privilégient la représentation du système des mots au détriment d'une forme schématique et argumentale propre à chaque unité. Inversement, les modèles qui cherchent à décrire cette forme schématique privilégient le contenu des unités au détriment de la représentation du système des analogies et des différences lexicales.

Un modèle synthétique devrait faire la somme des différentes caractéristiques. Au niveau global, il faudrait pouvoir distinguer les catégories mais aussi les valeurs sémantiques d'un mot et représenter les proximités et les différences. Au niveau de l'unité, il faudrait pouvoir construire un schéma décrivant la structure argumentale d'une unité et le mode de composition avec les autres unités d'un énoncé. Ce modèle devrait aussi permettre le passage et la cohérence entre ces deux niveaux d'organisation.

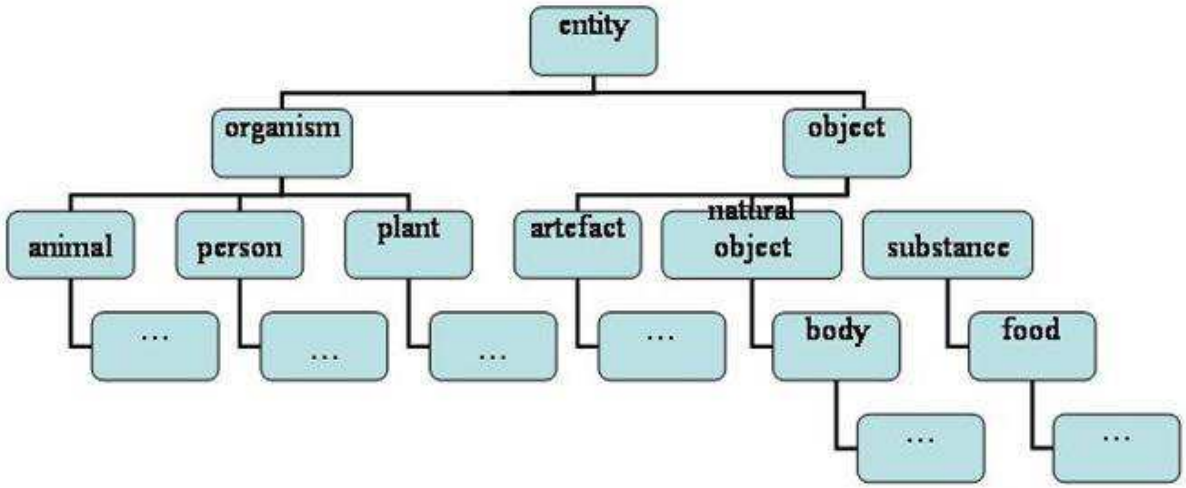

Figure 1 : Figuration d'une hiérarchie lexicale telle qu'elle est présentée par Miller [Fellbaum, 1998b].

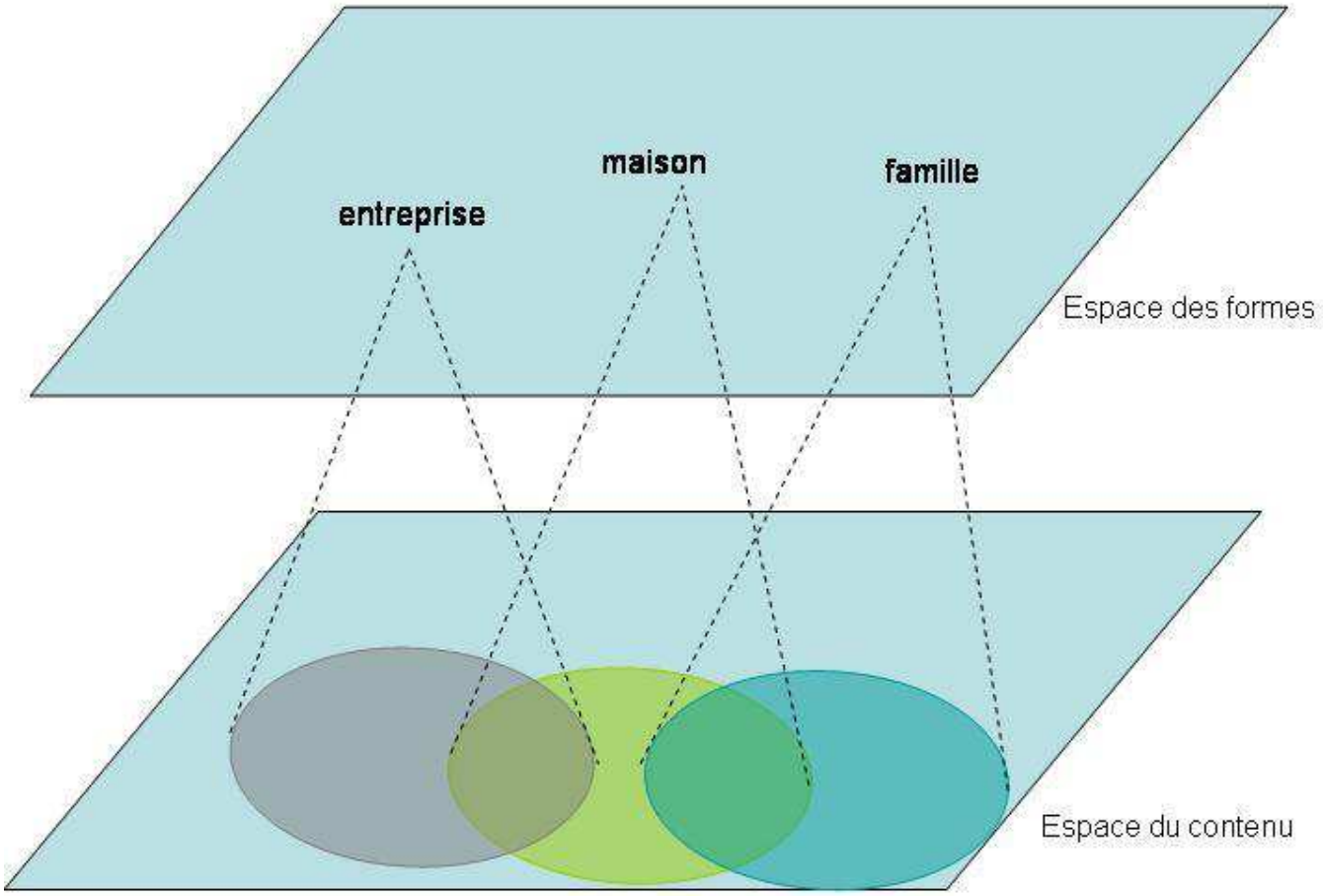

Figure 2 : Figuration du lien formes-contenu dans un modèle géométrique.

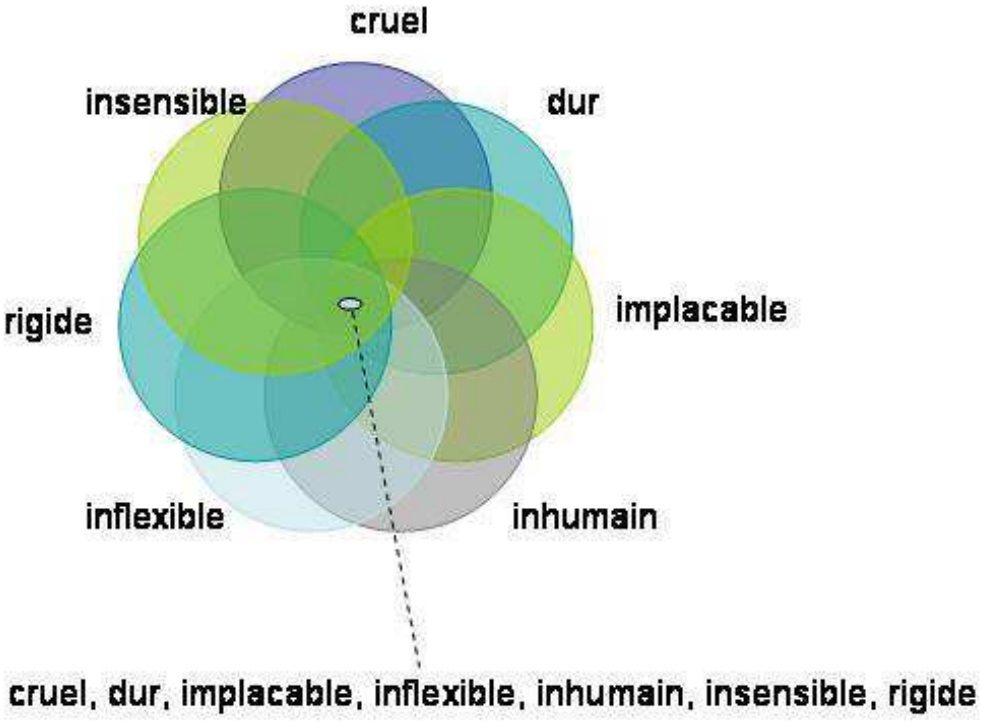

Figure 3 : Figuration de l'intersection des aires associées à des mots d'une même clique.

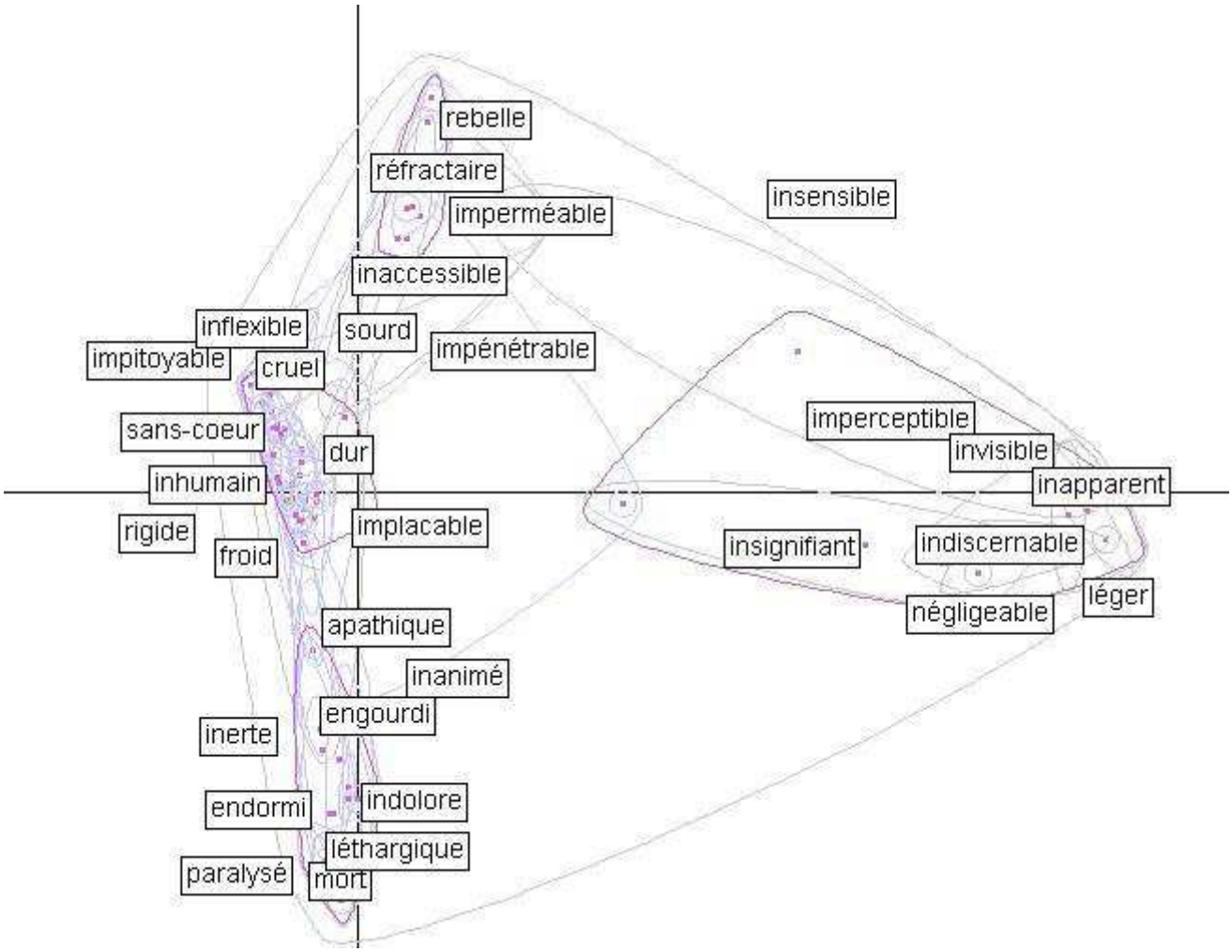

Figure 4 : Représentation géométrique associée au mot insensible, d'après [Ploux and Ji, 2003].

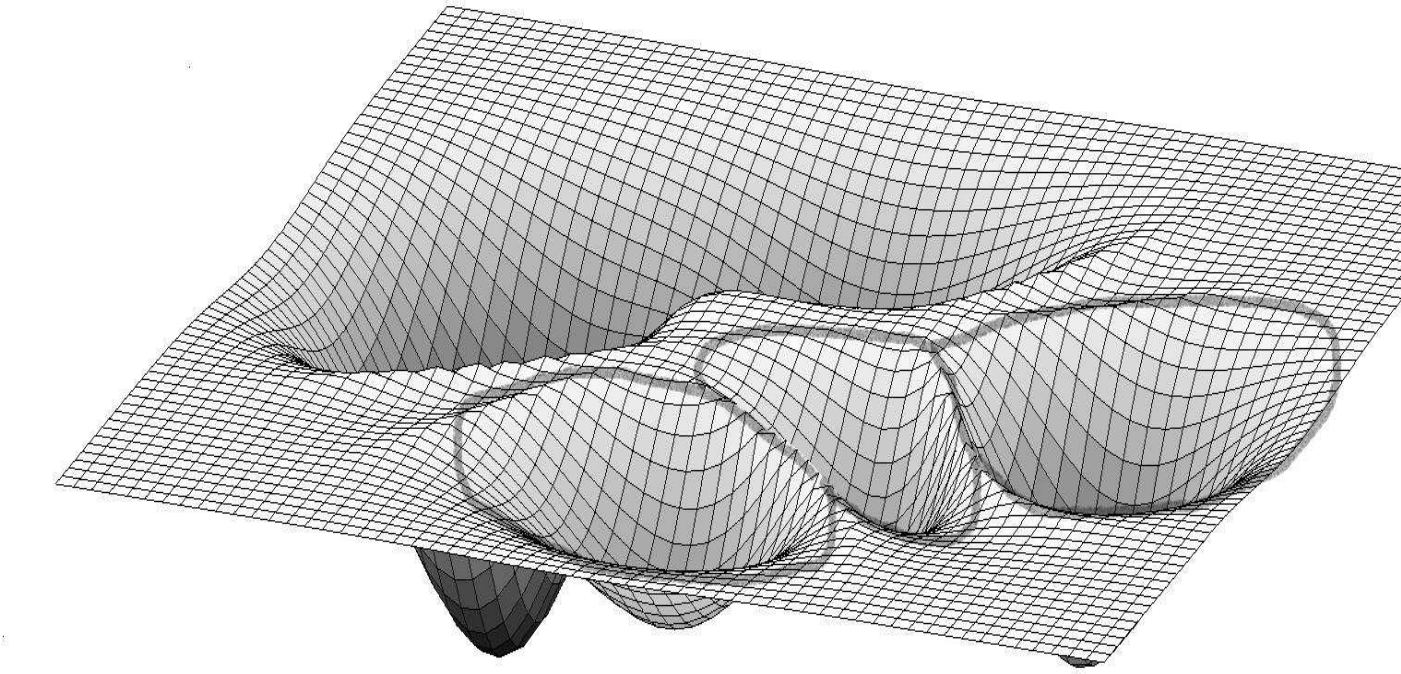

Figure 5 : Figure représentant des bassins d'attraction. Des lignes de crêtes séparant trois des bassins ont été ajoutées au trait.

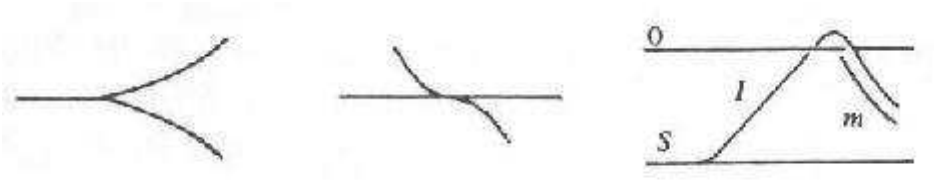

Figure 6 : Morphologies archétypiques des verbes *séparer, traverser, couper* (S pour sujet, O pour objet, I pour instrument, m représente la part de l'objet qui en a été détachée). Extrait de [Thom, 1977].

Footnotes:

[1]Une analyse factorielle est une méthode statistique d'analyse des données, [Bouroche and Saporta, 2002].
[2]Un formant est une valeur du spectre sonore, maximale en énergie.
[3]Une clique est un objet mathématique qui désigne un sous-graphe maximal, complet, connexe (il s'agit des plus grands sous-graphes possibles dont tous les sommets sont liés les uns les autres). Ici le graphe considéré est un ensemble de mots (les sommets) et de relations (ou arcs) qui lient ces mots.

Références

[Benzécri, 1980] Benzécri, J.-P. (1980). L'analyse des données : l'analyse des correspondances. Bordas, Paris.
[Bouroche and Saporta, 2002] Bouroche, J.-M. and Saporta, G. (2002). L'Analyse des données. Que sais-je ? PUF, 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.
[Chomsky, 1969] Chomsky, N. (1969). Structures syntaxiques. Paris, Seuil.
[Dowty et al., 1981] Dowty, D., Wall, R., and Peters, S. (1981). Introduction to Montague Semantics. D. Reidel Publishing Company, Dordrecht.
[Fairbanks and Grubb, 1961] Fairbanks, G. and Grubb, P. (1961). A psychophysical investigation on vowel formants. Journal of Speech and Hearing Research, 1 :203–219.
[Fellbaum, 1998a] Fellbaum, C. (1998a). A semantic Network of English Verbs, pages 23–46. MIT Press.
[Fellbaum, 1998b] Fellbaum, C., editor (1998b). Wordnet, An Electronic Lexical Database. MIT Press, Cambridge, Massachusetts.
[Forde and Humphreys, 2002] Forde, E. M. and Humphreys, G. W., editors (2002). Category-specificity in brain and mind. Psychology Press., East Sussex, UK.
[Gärdenfors, 2000] Gärdenfors, P. (2000). Conceptual Spaces, the Geometry of Thought. MIT Press, Cambridge, Massachusetts.
[Grefenstette, 1998] Grefenstette, G. (1998). Cross-language information retrieval, volume 2 of The Kluwer international series on information retrieval. Kluwer Academic, Boston, London.
[Habert et al., 1997] Habert, B., Nazarenko, A., and Salem, A. (1997). Les linguistiques de corpus. Armand Colin, Paris.
[Hauk et al., 2004] Hauk, O., Johnsrude, I., and F., P. (2004). Somatotopic representation of action words in human motor and premotor cortex. Neuron, 39(41) :301–307.
[Ji, 2004] Ji, H. (2004). A Computational Model for Word Sense Representation Using Contextual Relations. Mémoire de thèse en sciences cognitives.
[Ji et al., 2008] Ji, H., Lemaire, B., Choo, H., and Ploux, S. (2008). Testing the Cognitive Relevance of a Geometric Model on a Word-Association Task : A Comparison of Humans, ACOM, and LSA. Behavior Reseach Methods, 40(4) :926–934.
[Ji et al., 2003] Ji, H., Ploux, S., and Wehrli, E. (2003). Lexical knowledge representation with contexonyms. Proceedings of the 9th Machine Translation
Summit, pages 194–201.
[Kintsch, 2001] Kintsch, W. (2001). Predication. Cognitive Science, 25 :173– 202.
[Laham, 1997] Laham, D. (1997). Proceedings of the 19th annual meeting of the Cognitive Science Society, chapter Latent Semantic Analysis Approaches to Categorization, page 979. Mawhwah, NJ : Erlbaum.
[Landauer et al., 1998] Landauer, T. K., Foltz, P., and Laham, D. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, 25 :259– 284.
[Lund et al., 1996] Lund, K., Burgess, C., and Audet, C. (1996). Dissociating semantic and associative word relashionships using high-dimensional
semantic space. Cognitive Science Proceedings LEA, pages 603–608.
[Masson, 1995] Masson, M. (1995). A distributed memory model of semantic priming. Journal of Experimental Psychology : Learning, Memory, and Cognition, 21(1) :3–23.
[Miller, 1998] Miller, G. A. (1998). Nouns in WordNet, pages 23–46. MIT Press, Cambridge, Massachusetts.
[Moeschler, 2004] Moeschler, J. (2003-2004). Séminaire de pragmatique du discours. www.unige.ch/lettres/linge/ moeschler/ semantique2/sp2/sp2.ppt, Université de Genève.
[Moeschler and Auchlin, 2000] Moeschler, J. and Auchlin, A. (2000). Introduction à la linguistique contemporaine. Armand Colin, Paris.
[Montague, 1974] Montague, R. (1974). Formal Philosophy. Selected Papers of Richard Montague. Yale University Press, New Haven.

[Petitot-Cocorda, 1985] Petitot-Cocorda, J. (1985). Morphogenèse du sens. Presses Universitaires de France, Paris.
[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.
[Rastier, 1987] Rastier, F. (1987). Sémantique interprétative. PUF, Paris.
[Rosch, 1983] Rosch, E. (1983). New Trends in Cognitive Representation : Challenges to Piaget's Theory, chapter Prototype classification and logical classification : The two systems, pages 73–86. NJ : Lawrence Erlbaum Associates.
[Ruppenhofer et al., 2005] Ruppenhofer, J., Ellsworth, M., Petruck, M., and Johnson, C. (2005). FrameNet : Theory and Practice. http ://framenet. icsi.berkeley.edu/book/book.html.
[Smith et al., 1974] Smith, E. E., Shoben, E. J., and Rips, L. J. (1974). Structure and process in semantic memory : A featural model for semantic decisions. Psychological Review, 81(3) :214–241.
[Thom, 1977] Thom, R. (1977). Stabilité structurelle et morphogénèse. InterEditions, Paris.
[Thom, 1980] Thom, R. (1980). Modèles mathématiques de la morphogénèse. Christian Bourgeois Editeur, Paris.
[Victorri and Fuchs, 1996] Victorri, B. and Fuchs, C. (1996). Polysémie et construction dynamique du sens. Hermès, Paris.
[Vigliocco et al., 2004] Vigliocco, G., Vinson, D., Lewis, W., and Garrett, M.(2004). Representing the meanings of object and action words : The featural and unitary semantic system (fuss) hypothesis. Cognitive Psychology, 48 :422–488.
[Vossen, 2003] Vossen, P. (2003). The Oxford Handbook of Computational Linguistics, chapter Ontologies, pages 464–482. Oxford University Press.