ÉLÉMENTS POUR LA GÉNÉRATION DE CLASSES SÉMANTIQUES À PARTIR DE DÉFINITIONS LEXICOGRAPHIQUES. POUR UNE APPROCHE SÉMIQUE DU SENS

Mathieu VALETTE, Alexander ESTACIO-MORENO, Etienne PETITJEAN, Evelyne JACQUEY

ATILF (CNRS, Nancy)

(Article paru dans *Verbum ex machina*, *Actes de la 13ème conférence sur le traitement automatique des langues naturelles (TALN 06)*. Piet Mertens, Cédrick Fairon, Anne Dister, Patrick Watrin (éds). *Cahier du CENTAL*, 2.1, UCL Presses Universitaires de Louvain. Volume 1. Pages 357-366)

SOMMAIRE :

RÉSUMÉ : Ce papier expose une expérience de classification menée sur un corpus de définitions dictionnairiques. Le cadre général de cette recherche est la constitution d'une ressource lexico-sémantique fondée sur une conception structuraliste du sens (le contenu sémantique d'une unité lexicale est structuré en sèmes ; le sens d'un texte émerge de faisceaux de regroupements sémiques stabilisés). L'objectif de l'expérience rapportée est de découvrir des classes sémantiques à partir de définitions dictionnairiques avec la méthode CAH. Les classes sémantiques regroupent des unités lexicales en fonction de sèmes génériques (i.e. communs à toutes les unités lexicales de la classe) et s'organisent différentiellement en fonction de sèmes spécifiques. À partir d'une sélection d'entrées dictionnairiques partageant le sème générique /arbre/, nous étudions la distribution et l'organisation d'une hypothétique classe sémantique liée au domaine de la Sylviculture.

ABSTRACT: The paper exposes an experiment of classification based on a corpus of dictionary definitions. Underlying this research is the building of a lexico-semantic resource based on a structuralist approach to meaning (the semantic content of a lexical item is made up of semes; the meaning of a text emerges from groupings of stabilised seme sets). The purpose of the experiment is to make up semantic classes (or clusters) with dictionary definitions by using the HCA method. Semantic classes are built from lexical items according to generic semes (i.e., shared by all lexical items of the class). They are organised in a differential way according to their specific semes. From a selection of dictionary entries sharing generic seme /arbre/ ("tree"), we'll study the distribution and the organisation of an assumed semantic class linked to the domain of Forestry.

MOTS-CLÉS : ressources lexico-sémantiques, dictionnaire sémique, sémantique textuelle, classification automatique, CAH, Jaccard, lien moyen

KEYWORDS: lexico-semantic resources, seme dictionary, text semantics, clustering, HCA, Jaccard, UPGMA

1. Problématique générale

1.1. Brève typologie des ressources lexico-sémantiques actuelles

1.1.1. L'enrichissement de corpus en métadonnées constitue actuellement un champ de prospection important pour certaines applications linguistiques (fouille de textes, classification, filtrage) et pour l'ingénierie des connaissances (recherche d'information). La création de ressources lexico-sémantiques à cette fin est dominée depuis plusieurs années déjà par deux approches du sens, manifestement très productives à en juger par la quantité de ressources développées, mais souvent considérées comme encore insatisfaisantes (Slodzian 1999), (Véronis, 2004), (Rastier, 2001). La constitution de ces ressources relève de deux approches distinctes :

(i) une approche que nous qualifierons de *paradigmatique* (« en langue »), d'inspiration philosophique et terminologique, principalement représentée par les thésaurus et les ontologies (WordNet, les ontologies de domaines). L'approche paradigmatique propose une représentation close du monde ou du domaine, où la signification des items dépend de relations hiérarchiques (hyperonymie, hyponymie, etc.) construites en fonction des référents qu'ils désignent ;

(ii) une approche *syntagmatique* (« en discours »), d'inspiration logico-grammaticale et fondée sur le comportement cotextuel du lexique (FrameNet, VerbNet, Le Dico). Les unités lexicales y sont livrées accompagnées d'une notice d'actualisation décrivant la combinatoire syntaxique (arguments) et sémantique (actants) de leurs différentes acceptions.

On pourrait compléter cette brève typologie d'une observation transversale fondée sur le mode d'acquisition de la ressource. Nous distinguerons donc l'acquisition *empirique* où les ressources proviennent d'extraction sur corpus (FrameNet, terminologie textuelle) et l'acquisition *introspective* – à rapprocher de la *computer-aided armchair linguistics* telle que la qualifie plaisamment (Fillmore, 1992), où les ressources sont produites à partir de l'expertise du linguiste et éventuellement validées *a posteriori* par des collections d'attestations issues de corpus (WordNet, Le Dico).

1.1.2. Nous ne discuterons pas de la question de l'acquisition, l'intérêt de l'analyse en corpus étant reconnu depuis quelques années déjà, pour sa nature objectivante et sa robustesse (Fuchs & Habert, 2004), (Condamines, 2005). En revanche, il semble que les différentes ressources existantes, pour utiles qu'elles soient dans certains cadres applicatifs (comme par exemple le Web sémantique), n'en présentent pas moins quelques limites en termes de description des observables, notamment par rapport à notre objectif, l'*interprétation sémantique des textes assistée par ordinateur*.

Centrées sur la référence et non sur les usages en discours, les ontologies ne relèvent pas à proprement parler de la linguistique mais davantage de la philosophie. Légitime dans la perspective terminologique des ontologies de domaine (ou régionales), l'approche paradigmatique apparaît problématique dès qu'elle avoue une ambition intramondaine. Les relations entre items retenues par tradition s'avèrent, en effet, très insuffisantes pour rendre compte du sens des unités lexicales dès lors qu'elles sont actualisées dans des textes. L'hyperonymie et l'hyponymie en particulier, témoignent d'une représentation des connaissances ensembliste, abstraite, et non déterminée par les usages, lesquels sont fonctionnels par nature. Par exemple, *caviar*, dans l'ontologie WordNet, est en relation d'hyponymie avec *seafood* (« aliments issus de la mer »), au même titre donc, que le poisson pané, alors qu'il serait d'un usage plutôt antonymique et de toute façon bien davantage susceptible d'être actualisé en cooccurrence de *champagne* qui appartient à la

même classe sémantique des Mets festifs. Or, la subsomption ontologique de *caviar* et *champagne* est fort tardive parce que l'un est un solide et l'autre un liquide. Malgré une passerelle possible mais facultative au niveau d'un nœud *food* (*solid food* pour *caviar*, *nutrient* pour *champagne* – sic), ils sont séparés par l'abîme physique que constituent la pression et la température ambiante (*room temperature and pressure*). De fait, l'étude des cotextes de *caviar* dans les romans de FRANTEXT montre que le cooccurrent le plus fréquent de *caviar* est *champagne*, suivi de *foie gras* (fenêtre de 20 mots). On lira pour un développement (Rastier & Valette, à paraître).

Les approches syntagmatiques, plus pertinentes linguistiquement parlant parce qu'elles tiennent compte des conditions d'actualisation, reposent toutefois sur une vision grammaticale du sens, où le syntagme et la phrase (dont le parangon est la proposition logique) constituent les seules unités prises en compte. Or, si la signification d'une unité lexicale peut sans encombre être rapportée au lexique, son sens dépend dans une large mesure du texte dans son unité et du corpus dont celui-ci dépend, autrement dit, des *usages* socialement codifiés et linguistiquement organisés en discours (Rastier, 2001), (Véronis, 2004).

1.2. Pour une approche sémique du sens

1.2.1. Le cadre général de la recherche dans laquelle s'inscrit la présente étude est la réalisation d'une ressource lexico-sémantique alternative aux deux approches présentées ci-dessus. Du point de vue paradigmatique, nous proposons d'envisager les relations entre unités lexicales non pas en termes de construction hiérarchique mais en classes sémantiques. Du point de vue syntagmatique, l'instanciation des unités lexicales dans les textes ne se fait pas au niveau de la phrase ou de l'énoncé, mais au niveau d'unités textuelles plus larges (paragraphe, texte). Le cadre théorique choisi est inspiré de la sémantique textuelle (Rastier 2001) et induit un certain nombre de propositions liminaires :

a. Le contenu sémantique (sémie)[1] d'une unité lexicale (lexie) est constitué de traits sémantiques (sèmes).

b. Au sein d'une sémie, les sèmes sont organisés et pondérés en fonction des domaines, des discours et des genres textuels des différents corpus dans lesquels elle est actualisée, autrement dit, des usages possibles. Ainsi, à une lexie peut correspondre plusieurs configurations sémiques.
c. Les configurations d'une sémie sont obtenues par apprentissage. Les sèmes sont qualifiés en fonction de leur participation à des *réseaux sémiques intratextuels* dont on observe les régularités sur des corpus homogènes (en genre, discours et domaine). On appelle ces réseaux sémiques des *fonds* et des *formes sémantiques*. L'empan de ceux-ci varie suivant leur nature et celle du texte.

1.2.2. La ressource lexico-sémantique actuellement développée s'apparente donc à un *dictionnaire sémique* (appelé DIXEM), ou plus précisément, à une collection de dictionnaires sémiques relevant de domaines, de genres et de discours particuliers, destinée à l'analyse thématique et l'interprétation assistée. Cette recherche, qui s'inscrit dans la continuité des travaux actuels sur l'analyse thématique (Bourion, 2001), (Zweigenbaum & Habert, 2004), (Rossignol, 2005) mais dans une perspective infralexicale, prend partiellement appui, à la suite des travaux de (Martin, 2001) sur une exploitation du *Trésor de la Langue Française* (désormais *TLF*) et de *FRANTEXT*.

[1] Le métalangage employé dans cet article est dans une large mesure emprunté à la sémantique structurale, et particulièrement à la théorie développée par (Rastier, 2001). On se reportera à cet auteur pour un approfondissement.

1.3. Problématique de la présente étude

1.3.1. Si nous adoptons le point de vue développé ailleurs selon lequel la constitution d'une ressource lexico-sémantique implique de prendre en compte les usages (linguistiquement déterminés par les discours, les genres et les domaines), l'étude présentée ici vise à étudier, à titre préparatoire, les potentialités et les limites du recours à un dictionnaire de langue dans l'élaboration de ressources lexico-sémantiques. Elle est basée sur une interprétation dégradée de la notion de sème.

La question que nous nous posons est la suivante : la représentation sémique du contenu sémantique (sémie) permet-elle d'organiser le dictionnaire en classes sémantiques ? Rappelons que les classes, dans une perspective structurale, regroupent des unités lexicales (lexies) en fonction de sèmes génériques (i.e. communs à toutes les lexies de la classe) et se structurent en fonction de sèmes spécifiques (cf. Greimas, 1966, Rastier, 2001). Il s'agit donc d'observer : (1) si la définition lexicographique comprend les informations sémantiques suffisantes pour générer des classes sémantiques ; (2) si, le cas échéant, les classes sémantiques résultantes sont pertinentes, et de quel point de vue ; (3) si l'opposition structurale sèmes génériques/sèmes spécifiques y est observable et opératoire.

> Pour cette expérience, nous étudierons la distribution et l'organisation d'une hypothétique classe sémantique Sylviculture à partir d'une sélection d'entrées partageant le sème /arbre/, générique par hypothèse.

2. Description de l'étude

2.1. Constitution du corpus : autour du sème /arbre/

Le corpus est composé d'une sélection de définitions extraites du *TLF*. La première expérience de classification repose sur une hypothèse minimaliste : une définition est une sémie mise en texte. Ainsi, les mots pleins d'une définition (substantifs, adjectifs, verbes et certains adverbes) sont, une fois lemmatisés, les sèmes potentiels qui constituent la sémie d'une lexie en attente d'actualisation. Pour une définition telle que :

> *LAURACÉES. Famille de plantes dicotylédones, comprenant des arbres et des arbustes, à feuilles simples, alternes et persistantes, qui croissent dans les régions chaudes et tempérées.*

Nous extrayons la sémie suivante :

> *LAURACÉES {/famille/, /plante/, /dicotylédone/, /comprendre/, /arbre/, /arbuste/, /feuille/, /simple/, /alterne/, /persistant/, /croître/, /région/, /chaud/, /tempéré/}*

Cette sémie[2] vaut pour les besoins de l'expérience. Il va sans dire qu'un découpage par syntagme serait défendable, sinon exigible. Ainsi, des sèmes /région chaude/ et /région tempérée/ sembleraient plus pertinents que la triade obtenue. On comprendra par la suite que la brutalité de notre découpage est modérée par l'approche quantitative adoptée.

[2] Formellement, dans cette perspective lexicographique (et non textuelle), il s'agit d'un *sémème* et non d'une *sémie*, et d'un *lexème* et non d'une *lexie* ; mais pour la clarté de l'exposé, nous simplifierons le métalangage de la sémantique structurale en neutralisant ces distinctions émanant de l'opposition langue/discours.

Pour notre étude, nous avons donc extrait du *TLF* toutes les définitions contenant le mot *arbre*, de façon à obtenir un premier sous-corpus composé de 358 entrées (décrites par 6988 sèmes.

2.2. Distinguer les sous-classes de la classe hypothétique des arbres

2.2.1. Le premier volet de notre expérience vise à étudier, dans le corpus décrit ci-dessus, la valeur opératoire des sémies constituées à partir d'entrées, autrement dit, dans la perspective lexicographique adoptée, la pertinence de la représentation sémique pour la génération de classes sémantiques cohérentes.

Pour cela, nous avons procédé à une classification automatique, laquelle semble être la solution la plus naturelle. Cependant lorsqu'il s'agit de faire une classification automatique le choix est large. Deux grandes catégories de méthodes de classification automatique sont habituellement utilisées :

- Les méthodes à base de mesure de similarité qui regroupent des individus (entrées) similaires dans un même cluster (méthodes hiérarchiques (Lance & Williams, 1967), méthodes par partitionnement comme les k-moyennes (Duda & Hart, 1973), méthodes connexionnistes (Torres-Moreno *et al.*, 2000), etc.)
- Les méthodes probabilistes à base de mélange (Everitt & Hand, 1981) qui utilisent un mélange de densités pour modéliser un ensemble d'individus. Les classes correspondent aux différentes composantes du mélange.

2.2.2. Nous nous sommes orientés vers le premier type de méthodes et nous effectuons une Classification Ascendante Hiérarchique (CAH). La fréquence des sèmes par entrée étant en moyenne très basse, nous utilisons une représentation binaire des entrées[3].

Si V (vocabulaire) est l'espace des sèmes possibles, alors chaque entrée est représentée par un vecteur dans cet espace dont les composantes informent sur la présence (valeur égale à 1) ou l'absence (valeur égale à 0) d'un sème dans une entrée. De manière formelle, si E, est la représentation binaire d'une entrée de composantes E^l pour $l \in [1..|V|]$, où $|V|$ est la taille du vocabulaire, nous pouvons écrire :

$$\forall\, l \in [1..|V|], \quad E^l = \begin{cases} 1 & \text{si le } \textit{l-ème} \text{ sème apparaît dans } E \\ 0 & \text{sinon} \end{cases} \tag{1}$$

Puisque les sèmes sont des variables nominales asymétriques, c'est-à-dire que seulement les sèmes apparaissant dans la définition de chaque entrée sont intéressants pour calculer la distance entre deux entrées, nous utilisons comme mesure de dissimilarité le coefficient de Jaccard. Si l'on a deux entrées E_i et E_j, le coefficient de Jaccard, $d(E_i, E_j)$, se calcule ainsi :

$$d(E_i, E_j) = \frac{b + c}{a + b + c} \tag{2}$$

Où : a représente les sèmes partagés par les deux entrées, b les sèmes de l'entrée *i* n'apparaissant pas dans l'entrée *j* et c les sèmes de l'entrée *j* n'apparaissant pas dans l'entrée *i*.

[3] Précisons que pour cette première expérience tout du moins, nous avons neutralisé l'indication de domaine que propose (sporadiquement) le *TLF*.

Pour le rapprochement des classes nous utilisons la méthode du lien moyen (*average-linkage*). La distance entre deux classes C_K et C_M se calcule par l'équation :

$$D(C_K, C_M) = \frac{1}{N_K N_M} \sum_{i \in C_k} \sum_{j \in C_M} d(E_i, E_j) \quad (3)$$

Où : N_K et N_M sont les nombres d'entrées dans les classes C_K et C_M .

Finalement, nous avons exclu du tableau les cinq sèmes présentant les plus hautes fréquences. Parmi ceux-ci, on notera la présence logique de /arbre/ (477 occ.), qui de ce fait, subit un destin liposémique assez perecquien, et de quelques pseudo-sèmes bruyants tels que les auxiliaires /être/ (745 occ.) et /avoir/ (293 occ.). Le tableau soumis à la classification comprend donc 6983 sèmes.

2.2.3. Nous avons retenu 27 classes avec l'indice PSF (Calinski & Harabasz, 1974). Le Tableau 1 montre, pour les 5 classes les plus importantes, les effectifs (col. 2), la proportion d'entrées par classe (col. 3), le contenu global de chaque classe (col. 4) et quelques exemples d'entrées (col. 5).

Classe	Fréq.	Pourcent.	Contenu de la classe	Exemples
1	105	29,33	Essences d'arbre	*lauracées*, *rutacées*, *ramboutan*, etc.
2	94	26,26	Plantation (techniques et outils)	*racinage*, *scarifier*, *ombrophile*, etc.
3	17	4,75	Bûcheronnage	*lardoire*, *solivage*, *lambourde*, etc.
4	16	4,47	Parasites de l'arbre	*miastor*, *rhynchite*, *processionnaire*, etc.
5	15	4,19	Arboriculture	*sarter*, *palmette*, *provin*, *palissage*, etc.

Tableau 1. Caractérisation de la classification obtenue pour les cinq classes les plus importantes

L'analyse de la classification réalisée permet de distinguer, au sein de cette classe dont le seul élément structurant *a priori* est le sème /arbre/, plusieurs sous-classes sémantiquement homogènes. Si la granularité de l'analyse est insuffisante pour que des classes entièrement fonctionnelles émergent, l'on voit néanmoins se dessiner une typologie sommaire, avec des classes liées à des entités (essences, parasites) et des classes liées à des pratiques (arboriculture, bûcheronnage).

2.3. Caractérisation des sémies de la classe des Essences d'arbre

2.3.1. Le deuxième volet de notre expérience vise à approfondir notre connaissance d'une des classes dégagées ci-dessus, à savoir, celle des Essences d'arbre (la plus importante). Pour ce faire, nous avons choisi de pondérer toutes les sémies de façon à en dégager l'organisation interne relative à la classe considérée. L'objectif est de dégager, pour chaque sémie, les *sèmes génériques* (i.e. ceux susceptibles de structurer la classe ou des sous-classes) et des *sèmes spécifiques* (i.e. ceux qui décrivent de façon caractéristique chaque lexie) (cf. 1.3.1). Pour arriver à nos fins, nous utilisons un calcul classique en lexicométrie : l'écart réduit. Dans ce contexte l'écart réduit se calcule ainsi :

$$z = \frac{f_E - f_c * p}{\sqrt{f_c * p * q}} \quad (4)$$

Où f_E est la fréquence du sème observée dans l'entrée, f_c la fréquence du sème observée dans la classe, *p* est la proportion de l'entrée dans la classe et *q* le complément de *p*[4].

La pondération des sémies de chaque entrée à l'aide de l'écart réduit de la sous-classe des Essences d'arbre donne à voir des résultats globalement satisfaisants. Pour chaque sémie, les sèmes génériques (c'est-à-dire ceux dont l'écart réduit est le plus bas) permettent de distinguer des sous-classes potentielles. Par exemple, la lexie *sycomore* (tableau 3) semble appartenir à la sous-classe des essences qui valent pour leur bois et la lexie *monbin* (tableau 5) à celle valant pour leurs fruits.

Sème	Ecart réduit
jardin	20,84
avenue	20,84
ornement	10,35
légumineux	9,23
exotique	6,14
famille	1,84

Tableau 2 : Sémie pondérée de *sophora.* Définition : *Arbre exotique de la famille des Légumineuses, servant à l'ornement des jardins et des avenues.*

Sème	Ecart réduit
figuier	20,84
léger	11,98
imputrescible	11,98
érable	11,98
bois	3,09

Tableau 3 : Sémie pondérée de *sycomore.* Définition : 1. *Figuier au bois léger et imputrescible.* 2. *Érable.* 3. [P. méton.] *Bois de l'un de ces arbres.*

Sème	Ecart réduit
plage	18,04
lagune	18,04
maritime	18,04
intertropical	12,72
rhizophoracées	12,72
croître	4,03
région	2,88
famille	1,46

Tableau 4 : Sémie pondérée de *manglier.* Définition : *Arbre de la famille des Rhizophoracées, qui croît dans les lagunes et les plages maritimes des régions intertropicales.*

Sème	Ecart réduit
saveur	11,98
agréable	11,98
citron	11,98
pulpe	7,50
renfermer	6,82
taille	5,86
exotique	4,95
région	2,68
fruit	1,99

Tableau 5 : Sémie pondérée de *monbin.* Définition : *Arbre des régions exotiques, dont les fruits, de la taille d'un citron, renferment une pulpe à la saveur agréable.*

[4] Pour l'entrée *i* de la classe C_K, $p_{i,K} = \frac{\sum_{l=1}^{|V|} E_i^l}{N_K}$ et $q_{i,K} = 1 - p_{i,K}$

2.3.2. Cette classification interne à la sémie a évidemment une incidence déterminante sur les sèmes spécifiques, dans la mesure où ceux-ci sont spécifiques *par rapport* aux éléments de la la sous-classe esquissée par les sèmes génériques. Ainsi, /léger/ et /imputrescible/ caractérisent le bois du sycomore, tandis que /saveur/, agréable/, /citron/ renvoient au fruit du monbin. De la même façon, si /famille/ est le pseudo-sème générique commun à *sophora* (tableau 2) et *manglier* (tableau 4), leurs sèmes spécifiques caractériseront les arbres en tant qu'ils appartiennent à des familles (identifiables dans la sémies : légumineux dans un cas, rhizophoracées dans l'autre). Autrement dit, leur propriété remarquable est d'être des arbres et non de donner des fruits ou de produire un bois particulier. De fait, le sophora est caractérisé par sa fonction ornementale (sèmes spécifiques /jardin/, /avenue/, /ornement/) et le manglier par les endroits où il croît (sèmes spécifiques /plage/, /lagune/, /maritime/).

3. Discussion

Au vu des résultats obtenus, la réalisation préliminaire de classes sémantiques à partir de définitions dictionnairiques semble possible. Les items retenus qui composent une définition peuvent être légitimement considérés comme des traits sémantiques minimaux et ce, malgré une segmentation sommaire (et bruyante en conséquence) et la perte sensible d'information que l'éclatement syntaxique induit. La classification automatique permet de distinguer des sous-classes pertinentes, tant d'un point de vue gnosique (Essences d'arbre, Parasites) que praxique (Plantation, Bûcheronnage, Arboriculture) sans que de telles classes n'aient été dessinées *a priori*, ni par nous, ni par les lexicographes du *TLF*.

Par ailleurs, la caractérisation des sémies à l'intérieur d'une classe donne à voir une organisation sémantiquement pertinente, où les traits spécifiques sont susceptibles d'être opérationnels, notamment dans la perspective thématique/textuelle qui est la nôtre. En effet, les sèmes /jardin/, /avenue/ et /ornement/ semblent plus caractérisants du sophora que le fait qu'il s'agisse d'un arbre exotique, ou encore qu'il appartienne à la famille des légumineux (tableau 2). En d'autres termes, le *sens* apparaît valorisé au détriment de la *référence*. On imagine en particulier que les sèmes spécifiques participeront de façon privilégiée à des *formes sémantiques* (tandis que les sèmes génériques structureront le *fond sémantique*, cf. 1.2). Nous faisons l'hypothèse que l'apprentissage sur corpus, en faisant apparaître la régularité de ces réseaux de sèmes, permettra de stabiliser les sémies. Cette seconde phase du projet aura notamment pour effet de valider ou mettre à jour les résultats de notre classification préalable.

La nature continue des sémies obtenues par le calcul de l'écart réduit pose néanmoins la question du seuil. À partir de quelle valeur la sémie bascule-t-elle de la généricité à la spécificité ? Toutes nos tentatives visant à fixer automatiquement un seuil se sont avérées inappropriées. Les solutions restantes consisteraient, soit à fixer arbitrairement et prudemment *deux* seuils (l'un supérieur, l'autre inférieur) en renonçant à la qualification de l'entre-deux, ce qui reviendrait, par exemple, à allouer une valeur spécifique à /avenue/ et /jardin/, une valeur générique à /famille/ et à déqualifier /ornement/, /légumineux/ et /exotique/, ce qui est évidemment regrettable, notamment en ce qui concerne /ornement/ ; soit à faire l'hypothèse qu'une sémie est continue. Une nouvelle segmentation de la classe basée sur la variété des pseudo-sèmes génériques obtenus en 2.3 apporterait sans doute des éléments de solution ; nous laisserons néanmoins ouverte cette question qui trouvera peut-être un écho dans une certaine sémantique continuiste (Pottier 1992), (Visetti, 2004).

4. Ouverture

L'étude présentée ici avait, rappelons-le, une visée exploratoire. Il s'agissait de déterminer la pertinence d'un recours à un dictionnaire de langue dans l'élaboration d'une ressource lexico-sémantique (en l'occurrence, un dictionnaire de sèmes) destinée à l'annotation de

corpus, dans la perspective d'applications telles que l'analyse thématique et l'interprétation assistée. Les résultats se sont avérés agréablement positifs. Les biais lexicographiques attendus (artefactualité de la définition, décontextualisation méthodologique, etc.) se sont révélés moins déterminants que prévu. En somme, pour reprendre la distinction proposée en 1.1, il semble qu'une classification paradigmatique puisse participer, en amont et à titre préparatoire, à la réalisation d'une ressource lexico-sémantique de nature syntagmatique, laquelle ressortira à une linguistique d'aval, c'est-à-dire à une linguistique des textes.

BIBLIOGRAPHIE

BOURION E. (2001). *L'aide à l'interprétation des textes électroniques*. Thèse de doctorat, Université Nancy 2 ; publié sur *Texto ! Textes et cultures* (http://www.revue-texto.fr).

CALINSKI T., HARABASZ J. (1974). A Dendrite Method for Cluster Analysis. *Communications in Statistics*, 3, 1-27.

CONDAMINES A. (2005). Sémantique et corpus, quelles rencontres possibles ? *Sémantique et corpus*, A. Condamines, éd., Paris : Hermès.

DUDA R. O., HART, P. E. (1973). *Pattern Classification and Scene Analysis*. New York : John Wiley.

EVERITT B. S., HAND D. J. (1981). *Finite Mixture Distributions*. Monographs on Statistics and Applied Probability. New York : Chapman & Hall, London.

GREIMAS A. J. (1966). *Sémantique structurale*. Paris : PUF.

FILLMORE C.J. (1992). 'corpus linguistics or 'computer-aided armchair linguistics'. *Trends in Linguistics*, n°65, *Directions in Corpus Linguistics*, J. Svartvik, éd., Berlin : Mouton de Gruyter, 35-59.

FUCHS C., HABERT B., éd. (2004). *Traitement automatique et ressources numérisées pour le français*, *Le français moderne*, Vol. 72, n°1.

LANCE J., WILLIAMS W. (1967). *A General Theory of Classificatory Sorting Strategies*. Comput. J, n°9, 51-60.

MARTIN R. (2001). *Sémantique et automate*. Paris : PUF.

POTTIER B. (1992). *Sémantique générale.* Paris : PUF.

RASTIER F. (2001). *Arts et sciences du texte*. Paris : PUF.

RASTIER F., VALETTE M., (à paraître). De la polysémie à la néosémie. *Langages*, P. Siblot, éd., Paris : Larousse.

ROSSIGNOL M. (2005). *Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle*. Thèse de l'Université de Rennes 1.

SLODZIAN M. (1999). WordNet et EuroWordNet – Questions impertinentes sur leur pertinence linguistique. *Sémiotique*s, n°17, 51-70.

TORRES-MORENO J.M, VELAZQUEZ-MORALES P., MEUNIE J.G, (2000). Classphères : un réseau incrémental pour l'apprentissage non supervisé appliqué à la classification de textes. Actes des *JADT 2000*, pp 365-372, M. Rajman & J.-C. Chappelier éditeurs, EPFL.

VERONIS J. (2004). Quels dictionnaires pour l'étiquetage sémantique ? *Le Français moderne*, Vol. 72, n°1, *Traitement automatique et ressources pour le français*, C. Fuchs, B. Habert, éd., 27-38.

VISETTI Y.-M. (2004). Le continu en sémantique : une question de formes. *Cahiers de praxématique,* n°42, *Du continu : son et sens*, D. Ablali, M. Valette, éd., 39-74.

ZWEIGENBAUM P., HABERT B. (2004). Accès mesurés aux sens. *Mots. Les langages du politique*, n°74, 93-106.

Vous pouvez adresser vos commentaires et suggestions à : mvalette@atilf.fr

Référence bibliographique : VALETTE, Mathieu, et al.. Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens. Texto! [en ligne], mars 2006, vol. XI, n°1. Disponible sur : <http://www.revue-texto.net/Corpus/Publications/Valette_Estacio.pdf>. (Consultée le ...).