CHAPITRE III

# LES RESSOURCES LEXICALES POUR L'ETIQUETAGE SEMANTIQUE

Après la constitution de corpus de plus en plus volumineux, l'apparition de corpus étiquetés puis arborés, on commence à voir émerger des corpus porteurs d'annotations sémantiques. C'est un niveau d'annotation supplémentaire qui ouvre de nouvelles perspectives dans l'exploitation des corpus.

À l'heure actuelle, ces corpus porteurs d'annotations sémantiques n'existent cependant qu'à l'état embryonnaire[1]. Les expériences menées sont très diverses, reflets de conceptions sémantiques très différentes. L'essor des corpus arborés a fait suite à celui des corpus étiquetés et on peut s'attendre dans les prochaines années à l'apparition et au développement des corpus porteurs d'annotations sémantiques. Mais l'étiquetage sémantique est d'abord conditionné par la mise à disposition des connaissances sémantiques. La nature même des sources lexicales utilisées détermine en grande partie la méthode d'étiquetage et le jeu d'étiquettes retenus. Aujourd'hui, c'est donc la question de ces ressources qui paraît centrale.

Ce chapitre décrit les principales sources actuellement utilisées ou utilisables pour étiqueter sémantiquement des corpus. Seules les connaissances sémantiques sont prises en compte[2]. L'objectif est non pas de dresser un catalogue de ces ressources[3] mais d'en esquisser une typologie. Ces ressources ont été conçues selon des principes et dans des perspectives variées. Elles portent l'empreinte de ces différences de

[1] Ils ne dépassent guère 200 000 mots.

[2] Nous ne mentionnons donc pas les autres types de connaissances (phonétique, morpho-syntaxique…) que ces sources, les dictionnaires notamment, peuvent comporter.

[3] On trouvera ce type de catalogue sur des pages web régulièrement mises à jour. Un groupe de travail de l'Association for Computational Linguistics (ACL SIGLEX, Special Interest Group on the Lexicon) se charge notamment de recenser les ressources lexicales disponibles (http://www.clres.com/dict.html).

conception. Il s'agit ici d'évaluer dans quelle mesure elles peuvent servir à l'étiquetage sémantique de corpus et plus précisément à la désambiguïsation lexicale, même si ce n'est pas dans ce but qu'elles ont été conçues.

Les ressources sont donc considérées comme des bases de connaissances pour l'étiquetage sémantique des corpus (section 1). Elles sont de types variés. Elles diffèrent d'abord dans leur objet même, les unes portant sur des mots, les autres sur des notions ou concepts (section 2). La section 3 montre que ces bases de connaissances diffèrent également par la granularité de la description qu'elles donnent des mots, par leur degré de généralité et par leur codage. La section 4 présente ***WordNet***, l'une des sources lexicales les plus utilisées et le ferment de nombreux travaux de sémantique à partir de corpus. Nous terminons en soulignant le problème de la disponibilité des sources (section 5).

## 1. UN OBJECTIF: LA DESAMBIGUISATION LEXICALE

L'étiquetage sémantique consiste à attacher aux unités d'un texte (le morphème, le mot, une expression, un syntagme…) une étiquette sémantique qui indique selon les cas le sens du mot ou de l'expression, des traits ou catégories sémantiques, un marqueur de domaine ou de registre, etc.

À titre d'illustration, voici deux versions étiquetées d'une réponse de ***Enfants*** :

> je[sens=1] ne sais [sens=I.A.1] pas [sens=II.2], les [sens=I.1] gens [sens=I.A.1] sont [sens=II] égoïstes [sens=0] peut-être [sens=1].
>
> je ne[modalité=négative] sais [modalité=épistémique] pas [modalité=négative], les gens sont égoïstes peut-être [modalité=potentielle].

Dans la première version, à chaque mot est associée une étiquette reflétant le sens dans lequel il est employé : la distinction et la numérotation des sens est reprise du *Petit Robert*[4]. Dans ce cas, chaque mot est étiqueté[5]. Dans la deuxième version, en revanche, il s'agit d'un étiquetage partiel, qui ne concerne que les marques de modalités et qui devrait permettre d'observer la répartition de ces modalités dans l'ensemble du corpus. Comme au niveau syntaxique, ces étiquettes pourraient être complexes et combiner plusieurs traits.

Nous ne prenons ici en compte que le premier type d'étiquetage qui associe un ou plusieurs sens à un mot ou à une unité textuelle. On parle

[4] Dans l'édition de 1973. La valeur 0 indique que le mot a un sens unique.

[5] *Ne* ne porte pas d'étiquette sémantique parce qu'il n'a pas un fonctionnement autonome. Il forme avec *pas* un seul et même constituant discontinu.

dans ce cas de *désambiguïsation lexicale*[6] (*word sense disambiguation*). Il faut entendre ce terme dans un sens technique. L'objectif est d'identifier le *sens* dans lequel un mot est employé. Concrètement, il s'agit en fait d'un numéro de sens, ce sens étant choisi dans une liste finie de sens, laquelle est généralement issue d'une source de connaissances choisie comme référence (un dictionnaire, ici). La désambiguïsation est dite totale ou complète si à chaque mot est associé un sens et un seul. C'est le cas de l'exemple donné ci-dessus. On parle en revanche de désambiguïsation partielle si certains mots ne comporte pas d'étiquette de sens ou s'il en comporte plusieurs au contraire. Pour le verbe *sais* dans l'exemple ci-dessus, on aurait pu ainsi éviter de trancher entre différents sens très proches et laisser deux étiquettes : sais [sens=I.A.1] [sens=I.B.1]. Le degré de la désambiguïsation est une notion relative. D'un dictionnaire à l'autre les distinctions de sens ne se recouvrent pas : deux sens distingués dans l'un peuvent être confondus dans l'autre.

## 2. UNE OPPOSITION FONDAMENTALE : CONSTRUCTION LEXICALE OU CONCEPTUELLE

Une première distinction oppose les bases lexicales aux bases conceptuelles : les premières décrivent des mots et les secondes des objets[7] du monde tels que nous nous les représentons.

Mettons cette opposition en évidence à partir d'un exemple. Le mot *fauteuil* et la notion ou le *concept*[8] de fauteuil sont deux choses différentes. Le concept se définit traditionnellement soit par l'ensemble des chaises du monde réel auxquelles il renvoie, soit plutôt par un ensemble des propriétés[9] : un fauteuil est ainsi un siège comportant généralement quatre pieds, un dossier et des accoudoirs, un siège étant lui-même un meuble fait pour s'asseoir. Si le mot *fauteuil* se définit en partie comme le concept auquel il renvoie, il se définit aussi en opposition à tout un ensemble de mots comme *siège*, *chaise*, *tabouret*, *bergère*, par les connotations de confort, d'aisance et d'importance qu'il véhicule (« arriver dans un fauteuil », « fauteuil de président »), par ses emplois métonymiques (le fauteuil de président désignant souvent la fonction de président), etc.

Dans la pratique, les bases lexicales et conceptuelles dessinent deux espaces différents. Leur structure est parfois similaire : la relation SORTE-DE (IS-A, en anglais) de l'Intelligence Artificielle et de ses réseaux

[6] Lorsque le contexte est clair, nous parlons plus simplement de *désambiguïsation*.

[7] « Objet » est ici à entendre dans un sens large : il s'agit aussi bien d'objets concrets que d'entités abstraites ou d'événements.

[8] Nous ne parlons pas ici de notion mais de concept. Ce terme est utilisé en l'Intelligence Articificelle pour désigner l'image mentale que nous nous faisons des entités du monde, sans préjuger de la nature de cette image ou de son rapport au monde « réel ».

[9] On oppose ainsi les définitions extensionnelles et intentionnelles.

sémantiques est le pendant conceptuel de la relation d'hyponymie entre les mots[10]. L'opposition est parfois difficile à caractériser : on voudrait distinguer des catégories conceptuelles universelles ou du moins indépendantes de la langue mais force est de constater qu'un francophone et un anglophone — sans parler des inuits ou des mandchous — ne se représentent pas le monde de la même manière. Il reste que les bases lexicales et conceptuelles diffèrent dans leur visée : les unes décrivent le lexique ; les autres cherchent à modéliser le monde ou la représentation que nous nous en faisons. Les bases lexicales sont parfois utilisées pour construire des catégories sémantiques, et les bases conceptuelles pour décrire les mots, mais dans chaque cas ce n'est pas leur visée première.

## *2.1 Bases de connaissances lexicales*

La lexicographie cherche à recenser les mots d'une langue donnée et à les décrire, dans leurs différents sens, leurs relations et leurs emplois. Cette description peut se présenter sous différentes formes. De manière classique, nous distinguons les *dictionnaires*, les *thesaurus*, et les *terminologies*.

### 2.1.1 Dictionnaires

Les dictionnaires, qu'ils se présentent sous forme papier, sur support électronique ou qu'ils soient conçus pour le support électronique, qu'ils soient spécialisés ou de langue générale, contiennent les mêmes types d'informations sémantiques. La figure 3.1 ci-dessous en donne un exemple, tiré d'un dictionnaire électronique anglais[11].

Pour une langue donnée, les dictionnaires recensent les mots et les expressions considérées comme lexicalisées et donnent pour chacun une liste de sens, organisée en une arborescence de sens et de sous-sens. Chaque sens est décrit par une combinaison d'indications généralement optionnelles : une définition, un trait de domaine, des indications concernant le niveau de langue ou la modernité du mot, une liste de synonymes ou de renvois analogiques, des antonymes, des expressions ou tournures dans lesquelles entre le mot vedette, des phrases ou citations comme exemples d'emploi, ou même une ou plusieurs traductions possibles dans une autre langue[12]. La liste des sens pour un

[10] Cf. (Kleiber &Tamba, 1990).

[11] Nous donnons un exemple en anglais pour permettre la comparaison des informations données par les différentes ressources lexicales que nous évoquons dans ce chapitre, certaines de ces ressources (WordNet, en particulier) n'étant disponible que pour l'anglais. On pourra comparer cette entrée avec celle d'un dictionnaire français traditionnel donnée au chapitre VII, section 5.

[12] Les dictionnaires bilingues entrent en effet dans cette liste.

mot donné varie d'un dictionnaire à l'autre, leur description aussi. On a souvent souligné le nombre des définitions circulaires où deux ou plusieurs mots se définissent les uns par les autres, ainsi que le manque de cohérence dans la forme même des définitions ou l'ordre des indications. Il faut rappeler par ailleurs que les dictionnaires sont destinés à des locuteurs ayant déjà une bonne maîtrise de la langue dont ils ne fournissent qu'une description parcellaire. Il sont donc *a priori* peu adaptés aux traitements automatiques.

Pourtant, diverses expériences ont pris les dictionnaires comme sources de connaissances pour étiqueter les sens de mots, c'est-à-dire pour désambiguïser lexicalement les corpus. Il s'agit alors d'exploiter leurs distinctions de sens, chaque sens étant représenté, selon les cas, par sa définition elle-même et la liste des mots qu'elle contient (Véronis et Ide, 1990), par une mention de domaine (Guthrie *et al.*, 1991), par les différentes traductions possibles dans une langue cible, etc.

Après avoir dressé un panorama des travaux de désambiguïsation lexicale qui visent à assigner un sens aux mots d'un corpus, L. Guthrie *et al*. (1994, p. 87) reconnaissent que « [p]our le moment, beaucoup de chercheurs ont trouvé qu'un dictionnaire standard, avec ses distinctions de sens faites par des lexicographes professionnels, est la meilleure source de connaissances à exploiter pour la désambiguïsation. » En effet, les dictionnaires ont le mérite de proposer une description fine et relativement homogène de l'ensemble des mots courants. Les dictionnaires les plus complets décrivent les sens archaïques et rares, peu utiles pour le traitement des textes tout-venant, mais les dictionnaires usuels donnent une bonne description de la langue courante, même si certains sens dérivés et métaphoriques faciles à restituer par un être humain ne sont pas mentionnés.

[1]cred·it
Pronunciation: 'kre-dit
Function: *noun*
Etymology: Middle French, from Old Italian *credito*, from Latin *creditum* something entrusted to another, loan, from neuter of *creditus*, past participle of *credere* to believe, entrust -- more at CREED
Date: 1537
1 : reliance on the truth or reality of something <gave *credit* to everything he said>
2 a : the balance in a person's favor in an account b : an amount or sum placed at a person's disposal by a bank c : time given for payment for goods or services sold on trust <long-term *credit*> d (1) : an entry on the right-hand side of an account constituting an addition to a revenue, net worth, or liability account (2) : a deduction from an expense or asset account e : any one of or the sum of the items entered on the right-hand side of an account f : a deduction from an amount otherwise due
3 a : influence or power derived from enjoying the confidence of another or others b : good name : ESTEEM; *also* : financial or commercial trustworthiness
4 *archaic* : CREDIBILITY
5 : a source of honor <a *credit* to the school>
6 a : something that gains or adds to reputation or esteem : HONOR <took no *credit* for his kindly act> b : RECOGNITION, ACKNOWLEDGMENT <quite willing to accept undeserved *credit*>
7 : recognition by name of a person contributing to a performance (as a film or telecast) <the opening *credits*>
8 a : recognition by a school or college that a student has fulfilled a requirement leading to a degree b : CREDIT HOUR
synonym see BELIEF, INFLUENCE

*Figure 3.1.— Exemple d'entrée de dictionnaire : le nom credit*[13]

### 2.1.2 Thesaurus

Les thesaurus constituent un deuxième type de base de connaissances lexicales[14]. Ils organisent la description des sens de mots de manière différente des dictionnaires de langue. Ces derniers proposent avant tout des définitions de mots alors que les thesaurus reposent sur une sémantique plus spécifiquement relationnelle et servent à « mettre une idée en mots » ou à « trouver le mot juste ».

Les thesaurus comporte généralement deux voies d'accès. Un accès par les mots : comme les dictionnaires, les thesaurus comportent des entrées.

[13] Cet exemple est emprunté au dictionnaire de Merriam-Webster dans sa version en ligne : ***WWWebster Dictionary***, 1997, http://www.m-w.com/dictionary.htm (sept. 1997). La présence de mots en majuscules indiquant des renvois constitue la seule particularité de ce dictionnaire électronique : dans la version en ligne, il suffit de « cliquer » sur le mot CREED, pour en consulter l'entrée.

[14] Soulignons la différence des traditions lexicographique anglophone et francophone à cet égard : les anglo-saxons font grand usage de thesaurus mais c'est un outil méconnu des francophones. À l'inverse, ces derniers utilisent davantage les dictionnaires de langue.

Mais aussi un accès par les idées ou notions : les thesaurus regroupent les sens de mots en grandes catégories sémantiques et s'apparentent en cela aux ressources conceptuelles. Les figures 3.2 et 3.3 illustrent ces deux aspects.

La figure 3.2 montre qu'un mot, avec ses différents sens répertoriés, se définit par la place qu'il occupe dans un vaste réseau de mots et de sens, c'est-à-dire par les liens qu'ils entretient avec d'autres mots. Le thesaurus distingue quatre sens différents pour le nom *credit*, et pour chacun met lui associe des synonymes, des mots voisins, des antonymes et des mots opposés. L'exemple le montre, la définition quand elle est présente ne sert qu'à faciliter l'identification du sens.

credit
Function: *n*
Text: 1
Synonyms BELIEF 1, credence, faith
Related Word confidence, reliance, trust
2
Synonyms INFLUENCE 1, authority, prestige, weight
Related Word fame, renown, reputation, repute
Contrasted Words disrepute, ignominy, obloquy, opprobrium
Antonyms discredit
3 one that enhances another <he is a *credit* to his family>
Synonyms asset
Related Word honor
4 favorable notice or attention resulting from an action or achievement <took all the *credit* for the idea>
Synonyms acknowledgment, recognition
Related Word attention, notice; distinction, fame, honor; glory, kudos

*Figure 3.2.— Exemple d'entrée de thesaurus : le nom* credit[15]

Les thesaurus fournissent en fait un matériau plus directement utilisable que les dictionnaires pour la désambiguïsation lexicale. Ils donnent directement les associations de mots (synonymie, hyponymie, antonymies…) que l'on cherche à extraire, par divers traitements, des définitions de dictionnaire. Ils relèvent d'une vision relationnelle de la sémantique, proche de la conception distributionnelle qui sous-tend la plupart des travaux sur corpus (cf. chapitre VIII, section 5).

La structuration en catégories sémantiques est également exploitée pour l'annotation de corpus. Dans le ***Roget's Thesaurus***[16], plus de 30 000 mots sont réparties dans 1 000 catégories sémantiques (numérotées de #1 à #1 000), elles-mêmes organisées en cinq hiérarchies

[15] Cet exemple est emprunté au thesaurus de Merriam-Webster dans sa version en ligne : ***WWWebster Thesaurus***, 1997, http://www.m-w.com/thesaurus.htm (sept. 1997).

[16] Il s'agit du ***Roget's Thesaurus*** de 1911 dans sa version électronique, actuellement disponible à l'adresse http://ecco.bsee.swin.edu.au/text/roget/headings.html.

de faible profondeur (cinq niveaux au maximum) (cf. figure 3.3). On voit donc apparaître deux niveaux possibles de catégorisation : aux feuilles de la hiérarchie des regroupements lexicaux ; dans la structure, une catégorisation conceptuelle.

De fait, diverses expériences[17] ont montré l'intérêt que présentent les catégories sémantiques d'un thesaurus comme le ***Roget's*** pour la désambiguïsation lexicale.

Class I : Words Expressing Abstract Relations
SECTION I. EXISTENCE
1. BEING, IN THE ABSTRACT
#1. Existence.
#2. Inexistence.
…
SECTION II. RELATION
…
Class V : Words Relating to the Voluntary Powers
DIVISION (1) INDIVIDUAL VOLITION
SECTION I. VOLITION IN GENERAL
1. ACTS OF VOLITION
#600. Will.
#601. Necessity.
…
Class VI : Words Relating to the Sentient and Moral Powers
…
#998. Rite.
#999. Canonicals.
#1000. Temple.

*Figure 3.3.— Organisation générale des 1 000 catégories conceptuelles du* ***Roget's Theasaurus***

### 2.1.3 Terminologies

Les terminologies constituent un troisième type de ressources lexicales. Généralement établies pour des domaines spécialisés, elles sont peu adaptées à la désambiguïsation de vastes corpus. Outils traditionnels de la recherche documentaire (cf. chapitre IV, section 3), elles visent à recenser les dénominations d'un domaine (cf. chapitre II, section 3.4) et peuvent également servir à marquer les termes dans le cadre d'un étiquetage partiel de corpus.

[17] Voir notamment (Grefenstette, 1996) ou (Yarowsky, 1992).

### *2.2 Bases de connaissances conceptuelles*

Alors que les ressources lexicales structurent l'espace des mots, les *réseaux sémantiques* et *ontologies*, issus d'une autre tradition aussi ancienne que la lexicographie[18], reflètent une conceptualisation du monde. Il s'agit cette fois de recenser les « catégories d'objets » ou concepts du domaine considéré et éventuellement de représenter leurs propriétés ainsi que les relations qu'ils entretiennent entre eux. Il en résulte des hiérarchies ou des réseaux de concepts.

Les ontologies proposent un découpage du monde — ou de la représentation que nous en avons — en catégories, ces catégories étant organisées en hiérarchie par des liens SORTE-DE (*IS-A*). Lorsque s'y ajoutent d'autres types de relations (relations de causalité, d'appartenance, etc.) on obtient non plus un arbre ou une hiérarchie mais un graphe, un « réseau sémantique » ou « conceptuel » dans la terminologie de l'Intelligence Artificielle.

Initialement cantonnés à des domaines très spécialisés ou à des exemples de taille limitée, ces réseaux servaient surtout à valider une approche, un formalisme ou une théorie. La décennie présente voit cependant apparaître des bases de connaissances conceptuelles de grande ampleur. Le projet ***Cyc*** est exemplaire à cet égard (Guha et Lenat, 1990). Commencée il y a plus de 10 ans, l'ontologie, pièce centrale de cette base de connaissances contient aujourd'hui des dizaines de milliers de nœuds ou concepts. Pour ses concepteurs, le haut de cette hiérarchie qui comporte plus de 3 000 concepts est formé de catégories universelles.

### *2.3 Une opposition réelle mais floue*

Les ressources conceptuelles ont l'avantage de s'affranchir du niveau de structuration proprement lexical qui regroupe les différents sens d'un mot polysémique et qui représente les synonymes par des unités distinctes même si elles sont sémantiquement liées. Le mode de structuration conceptuel est plus proche du sens des mots que des mots eux-mêmes et donc mieux adapté à l'objectif de la désambiguïsation lexicale.

À l'inverse, quand il s'agit d'étiqueter un corpus, on a affaire à des mots. Établir le lien entre un concept ou une primitive ontologique et ses réalisations linguistiques, l'ensemble des mots qui y renvoient, ne va pas de soi. L'expérience de modélisation du projet Menelas (Zweigenbaum, 1994) a mis en évidence la nécessité de construire un lexique sémantique, interface entre une ontologie, objet conceptuel, et le texte, pour faire le lien entre le concept et le mot. De la même manière, les concepteurs de l'ontologie ***Cyc*** prévoient une interface linguistique.

---

[18] Cette tradition, qui remonte à la métaphysique antique, a été largement revisitée depuis une trentaine d'années par les recherches dans le domaine de l'Intelligence Artificielle.

L'opposition est cependant loin d'être nette. Les thesaurus, on l'a vu, sont des objets hybrides et les noms des classes supérieures de la hiérarchie du ***Roget's thesaurus*** : « *words expressing*… » (*mots exprimants*…) soulignent l'ambivalence conceptuelle et lexicale de cette hiérarchie. De fait, les mots ne s'organisent pas facilement en une hiérarchie bien structurée : le niveau supérieur, qui est abstrait et qui recouvre des grandes notions peu représentées dans le lexique, est généralement structuré *in abstracto* avec parfois de nouveaux concepts ou termes créés pour les besoins de la structuration.

À l'inverse, en dépit de l'ambition parfois affichée, il paraît illusoire de croire à l'universalité de l'ontologie résultante et de penser qu'une conceptualisation du monde puisse être indépendante de la langue de son concepteur. Concrètement, cette dépendance est en particulier marquée dans le fait que les nœuds et les relations d'un tel réseau conceptuel portent des étiquettes empruntées au langage naturel, ce qui conditionne et biaise l'interprétation.

## 3. UNE GRANDE DIVERSITE DE RESSOURCES LEXICALES

Au-delà de cette distinction entre ressources lexicales et ressources conceptuelles, différents paramètres sont à prendre en compte dans le choix d'une base de connaissances pour un projet donné.

### *3.1 Des distinctions de sens plus ou moins fines*

Les bases lexicales fournissent généralement des distinctions de sens fines. Le *Petit Robert*[19] liste douze sens pour le nom *cours*, répartis en 6 sens principaux. Le *Webster's Collegiate Dictionary*[20] distingue trois entrées pour le nom *bank* et au total seize sens différents. ***WordNet*** ou ***le Roget's thesaurus*** distinguent respectivement 8 et 20 acceptions pour le mot *credit.*

On peut rechercher au contraire des distinctions de sens plus grossières, ce qui réduit le nombre de sens et donc la polysémie des mots.

Les dictionnaires établissent des distinctions « homographiques »[21] (Guthrie *et al*., 1994), représentées soit par des entrées distinctes, soit par les premières divisions de sens. Ainsi, pour l'anglais *bank*, on peut différencier l'*établissement bancaire* et la *berge,* pour le français *cours*, on

[19] Dans l'édition de 1972.

[20] Dans la 9e édition.

[21] Il s'agit plutôt de grandes familles de sens que de « vrais » homographes, ces sens pouvant être dérivés les uns des autres.

peut distinguer les sens de *écoulement* et de *enseignement,* sans pour autant prendre en compte toute la diversité des sens donnés par les dictionnaires. Les dictionnaires donnent par ailleurs des distinctions de domaine (médecine, législation, technique…) qui sont elles aussi exploitables dans la perspective de la désambiguïsation lexicale (Guthrie *et al*., 1991).

Ces distinctions grossières peuvent également être obtenues à partir de thesaurus. Il faut alors tirer parti du haut de la hiérarchie des sens. Ces bases lexicales sont généralement structurées comme un ensemble de hiérarchies distinctes, chacune étant dominée par une catégorie sémantique générale. Pour un mot, on peut ainsi distinguer des grandes familles de sens sur la base de l'appartenance des sens à l'une ou l'autre de ces hiérarchies. C'est l'approche de R. Basili *et al*. (1997, p. 248) qui ne retiennent, pour travailler sur les verbes, que 15 grandes catégories de ***WordNet*** (perception, émotion, création, changement…) et ignorent les distinctions plus fines internes à chaque catégorie. Le verbe anglais *record* ou son équivalent français *enregistrer* admettent ainsi en langue générale, trois sens représentés par les catégories de la cognition, de la communication et de la perception. E. Agirre et G. Rigau (1996) exploitent de la même manière les 25 grandes catégories de noms de ***WordNet*** pour établir des grandes oppositions de sens. Dans (Bouaud *et al*., 1997), « une catégorisation à gros grain » est élaborée de la même manière à partir d'une nomenclature médicale dans la perspective d'un étiquetage sémantique de Menelas.

Si ces sources permettent de décrire des distinctions de sens fines ou grossières, il est généralement plus difficile d'établir des distinctions intermédiaires. Les distinctions et hiérarchies de sens des dictionnaires ou thesaurus ne reflètent pas une description homogène dans sa granularité. De fait, dans WordNet, certains liens hyponymiques reflètent une proximité sémantique beaucoup plus grande que d'autres : « [on trouve] des liens qui semblent représenter, pour certains, une courte distance (RABBIT-EARS IS-A TELEVISION-ANTENNA) et pour d'autres, une longue distance (PHYTOPLANKTON IS-A LIVING-THING) »[22] (Resnik, 1995a).

### *3.2 Des ressources générales ou spécialisées*

Il faut également distinguer les sources qui permettent de décrire la langue générale et celles qui rendent compte d'une langue spécialisée[23].

Les bases lexicales générales sont peu adaptées au traitement de corpus spécialisés : « nous avons montré que les sens de mots proposés par la plupart des dictionnaires électroniques accessibles en ligne ne

[22] Soit, littéralement : OREILLE-DE-LAPIN SORTE-DE ANTENNE-DE-TELEVISION et PHYTOPLANCTON SORTE-DE ETRE-VIVANT. En anglais, on appelle *rabbit ear* (*oreille de lapin*) les antennes de télévision en forme de « V ».

[23] Bien que des projets pour la construction d'ontologies générales existent (comme le projet ***Cyc*** mentionné ci-dessus), aucune expérience, à notre connaissance, n'a été faite pour utiliser ces ontologies pour le traitement de corpus.

permettent souvent pas d'exprimer les sens de mots dans un contexte spécifique. Certains emplois spécifiques (*i.e.*, techniques ou simplement jargonnants) sont souvent absents des sources à visée générale (comme ***WordNet*** ou le ***Longman Dictionary of Contemporary English***) […]. Ces sources sont donc trop peu spécifiques (en ce qui concerne le langage du domaine) et trop générales (parce qu'elles donnent une vue vague de la langue, indépendante de toute application). » (Basili *et al*., 1997, p. 237) Trop peu spécifiques dans la mesure où certains mots et certains sens de mots spécialisés ne sont pas représentés. Trop générales car elles décrivent la diversité des sens de la langue générale alors que la polysémie est souvent réduite dans les textes produits dans des domaines spécialisés.

Malheureusement, les sources spécialisées font souvent défaut et celles qui existent ne peuvent pas être réutilisées dans une perspective différente de celle pour laquelle elles ont été conçues initialement. L'expérience de (Charlet *et al*., 1996) est instructive à cet égard. Travaillant dans le domaine médical où les expériences de ce type sont anciennes, ces auteurs ont cherché, pour modéliser le domaine des maladies coronariennes, à réutiliser une base de connaissances préexistante, ***Unified Medical Language System*** (***UMLS***, (Humphrey et Lindberg, 1989)), précisément conçue comme un réseau sémantique unifié pouvant être utilisé dans différentes perspectives. Cette tentative s'est soldée par un échec et les deux principales raisons invoquées ne sont en rien spécifiques à cette expérience. La première concerne la couverture du domaine. Même si UMLS est une base de connaissances spécialisée, les auteurs font un constat similaire à celui que fait R. Basili pour les ressources lexicales générales : ils ont du enrichir certaines parties de la hiérarchie. La seconde est plus fondamentale : l'ontologie d'un domaine dépend d'un point de vue sur ce domaine et de la tâche qui est visée et de la tâche pour laquelle elle a été conçue ; elle n'est donc réutilisable que dans la mesure où la tâche demeure la même, ce qui est rare[24].

Les ressources lexicales font donc particulièrement défaut lorsqu'on se propose de traiter des corpus spécialisés. Deux autres pistes sont explorées. La première consiste à spécialiser une source lexicale générale pour l'ajuster à un domaine de spécialité. R. Basili et ses collègues tentent ainsi d'adapter la taxonomie des verbes de ***WordNet*** à divers domaines spécialisés en se fondant sur l'information contextuelle apportée par un corpus représentatif du domaine considéré. Ils distinguent les sens de verbes selon leur appartenance aux 15 grandes catégories sémantiques de ***WordNet*** (changement, cognition, communication, contact, émotion…). Il s'agit de sélectionner, parmi les différents sens associés à un verbe donné, ceux qui sont pertinents dans le domaine et d'ajouter les sens

[24] « [L]orsque les connaissances ont des dépendances par rapport à la tâche qui sont parfaitement connues et constantes, on peut faire des ontologies réutilisables ; pour Menelas c'est le cas des médicaments (et c'est le seul) : la description du Vidal (dictionnaire des médicaments) fournit toute les connaissances nécessaires pour prendre en compte tous les usages que l'on peut faire d'une ontologie des médicaments dans un cadre thérapeutique, et c'est ce cadre qui est sous-tendu par la plupart des applications médicales qui ont besoin d'une ontologie des médicaments. » (Charlet *et al.*, 1996).

spécialisés qui ne seraient pas représentés dans le réseau initial[25]. La seconde piste vise à constituer les ressources lexicales dont on a besoin. Cette construction peut être manuelle mais cela limite considérablement la finesse de la description. R. Basili *et al*. (1993a) décrivent une expérience de ce type : ils utilisent une quinzaine de catégories très générales (action, artefact, lieu, matière…) pour étiqueter des textes spécialisés. Elle peut également être automatique. Il s'agit alors d'acquérir des connaissances lexicales spécialisées à partir des corpus du domaine : de nombreux travaux se situent dans cette optique, nous y revenons au chapitre IV.

## *3.3 Des sources plus ou moins informatisées*

Les ressources utilisables se distinguent enfin par la forme sous laquelle elles se présentent. Entre les dictionnaires ou terminologies classiques sur support papier et un réseau sémantique doté d'une interface évoluée comme ***WordNet***, il y a divers degrés d'informatisation. Il va de soi qu'une ressource informatisée permet des traitements plus divers et à moindres coûts.

### 3.3.1 Dictionnaires et thesaurus sur support électronique

Les bases lexicales sur support électronique, les dictionnaires notamment (*machine-readable dictionaries*), se situent à un premier niveau. On désigne ainsi les versions électroniques des dictionnaires, thesaurus, terminologies et autres bases de connaissances disponibles qui ont été saisies ou scannées. Par rapport à la version « reliée », seul le support change : les données sont identiques. Pourtant ce premier niveau d'informatisation permet déjà de nouveaux modes d'exploration.

Dans un dictionnaire qui se présente sous la forme d'un livre, on ne peut guère rechercher les mots qu'au hasard ou par ordre alphabétique. C'est là la limite des dictionnaires traditionnels pour G. Miller, le « père » de ***WordNet*** (1993). Considérant un exemple de définition hyperonymique de *arbre* (*tree*) pris au sens de *plante*, il regrette qu'elle soit « terriblement incomplète » : le sens dans lequel l'hyperonyme *plante* doit être entendu n'est pas spécifié, on ne sait pas s'il existe d'autres plantes qui ne soient pas des arbres, on ne peut pas retrouver facilement les différentes sortes d'arbres.

Dès lors que le texte est sur support électronique, on peut facilement passer d'une entrée à l'autre ; par des algorithmes sur les chaînes de caractères, on peut trouver les mots ayant une terminaison commune,

[25] Leur démarche consiste à identifier pour chaque catégorie sémantique un noyau de verbes représentatifs et à repérer les contextes dans lesquels ces verbes figurent pour construire une description distributionnelle de chaque catégorie, puis à assigner un ou plusieurs sens à un verbe en comparant sa distribution avec celles des classes sémantiques.

rechercher tous les mots dont les définitions contiennent un mot donné, etc. Cela permet de s'affranchir partiellement des limites des définitions mentionnées par G. Miller : on peut reconstituer une partie de l'information manquante dans l'entrée de *arbre* en recherchant les entrées qui comportent les mots *arbre* ou *plante* dans leurs définitions.

### 3.3.2 Ressources électroniques

Dans les ressources qui ne constituent que les versions électroniques de dictionnaires traditionnels, cependant, l'information véhiculée par la typographie et la mise en page peut être difficile à exploiter, quand elle n'est pas purement et simplement perdue. Or elle est importante pour l'utilisateur : elle indique le statut des informations et guide l'interprétation de l'utilisateur. Pour préserver cette information et la rendre exploitable, il faut donc l'encoder. Nous revenons au chapitre VII sur les principes d'un tel encodage. L'important ici est de distinguer les ressources sur support électronique et les ressources électroniques en tant que telles, dont le codage est conçu pour faciliter l'accès par des traitements automatiques, pour expliciter le statut des informations données et donc en fournir les règles d'interprétation.

### 3.3.3 Ressources informatisées

La mise sur support informatique des ressources lexicales ouvre la voie à des nouveautés plus radicales.

S'affranchir du support papier, c'est d'abord s'affranchir de l'ordre linéaire. La structuration du dictionnaire en entrées distinctes, la numérotation des sens et les diverses marques typographiques étaient des premiers pas pour échapper à cette contrainte et donner un accès « direct » à certaines données. Pour autant, il n'était pas possible de consulter en parallèle plusieurs entrées d'un dictionnaire, de repérer des symétries, des parallélismes et plus généralement la structure sous-jacente à un ensemble de mots sans un long parcours de renvois en renvois et un patient travail de reconstitution. De la même manière, pour se faire une idée générale de la hiérarchie d'un thesaurus, il est important de pouvoir varier le niveau de description[26], une approche dynamique que ne permettait pas le support papier. L'outil informatique permet désormais de structurer les ressources lexicales sur d'autres bases et la multiplication des liens entre les différents éléments d'information autorise de nouveaux modes de consultation. ***WordNet*** en est un exemple intéressant (cf. section 4)[27].

---

[26] Soit en faisant un « zoom » pour concentrer son attention sur une zone donnée soit au contraire en faisant abstraction d'un certain niveau de détail pour dégager une vue d'ensemble.

[27] « En termes de couverture, les objectifs de ***WordNet*** diffèrent peu de ceux d'un bon dictionnaire standard de langue. C'est dans l'organisation de cette information que

En conséquence, les dictionnaires électroniques permettent de gagner en cohérence. Prenons pour seul exemple le travail effectué sur le français par I. Warnesson (1985) pour constituer, à partir de différentes sources traditionnelles, un nouveau dictionnaire des synonymes reposant sur une définition formelle de la synonymie comme relation d'équivalence[28]. La cohérence d'un tel dictionnaire en faciliter l'exploitation.

Dans ce domaine de la lexicographie, l'informatique a déjà induit de profonds bouleversements, avec notamment de nouveaux modes de navigation et de nouvelles possibilités d'exploration, mais il reste probablement à inventer de nouvelles formes de dictionnaires. On peut penser en particulier à des bases de connaissances intégrées et dynamiques, aux degrés de granularité et de spécialisation variables, qui puisse être reconfigurées en fonction des besoins et des parcours de l'utilisateur et offrir ainsi différents points de vue à l'utilisateur. Reprenons l'exemple de *credit*. C'est un mot polysémique, riche en connotations et son entrée dictionnairique est trop riche pour être facile à exploiter. Si l'utilisateur s'intéresse au domaine économique et financier, la plupart des sens deviennent immédiatement caduques tandis que les détails du deuxième sens prennent de l'importance. On devrait ainsi pouvoir considérer une base de connaissances sous différents points de vue.

## 4. UN EXEMPLE DE RESEAU LEXICAL : WORDNET

Nous présentons ici l'exemple de ***WordNet***, un thesaurus électronique. Deux raisons président au choix de cette base lexicale. C'est probablement la base de connaissances générales la plus utilisée : elle a servi à mettre au point ou à tester de nombreuses expériences depuis le début des années 1990. Par ailleurs, ***WordNet*** est un exemple d'une base lexicale conçue et pensée pour le support électronique.

### *4.1 Un projet ambitieux*

Depuis 1985, un groupe de psycholinguistes et de linguistes de l'université de Princeton a développé une base de données lexicale selon des principes suggérés par des expériences et des recherches en psycholinguistique sur l'organisation de la mémoire humaine. Depuis cette date, ce projet a pris de l'ampleur ; il se poursuit encore de nos jours. Le réseau ***WordNet*** disponible aujourd'hui est la version 1.5. Il peut soit être

---

***WordNet*** prétend innover. » (Miller *et al.*, 1993, p. 1).

[28] Qui respecte les propriétés de symétrie, de transitivité et de réflexivité.

consulté en ligne soit être importé[29].

### 4.1.1 Représenter les sens de mots

L'objectif de ***WordNet*** est de décrire comment les sens de mots ou concepts[30] — et non les mots eux-mêmes — s'organisent les uns par rapport aux autres. En ce sens, WordNet ressemble davantage à un thesaurus qu'à un dictionnaire. La théorie sous-jacente est une théorie différentielle[31] : un sens se définit par la place qu'il occupe dans le réseau, par les relations de proximité ou de contraste qu'il entretient avec les sens voisins. Partant de ce principe, un sens est représenté par un ensemble de synonymes : « Les ensembles de synonymes (*synsets*) n'expliquent pas ce que sont les concepts ; ils en posent l'existence. On suppose que les locuteurs anglais ont déjà acquis ces concepts et sont en mesure de les reconnaître à partir des mots listés dans le synset. » (Miller *et al*., 1993, p. 5-6). Considérons l'exemple du mot *credit* pour lequel huit sens sont identifiés dans ***WordNet***[32]. En voici trois:

1. credit (money available for a client to borrow)
2. recognition, credit (approval ; « give her recognition for trying » ; « he was given credit for his work » ; « it is to her credit that she tried »)
3. credit, deferred payment (arrangement for deferred payment for goods and services)

À chaque sens sont associés des synonymes, dans la mesure où il en existe. Parler du deuxième sens de *credit* ou du synset {*recognition*, *credit*} revient au même. Les définitions ou exemples (notés entre parenthèses) qui sont souvent associés aux concepts dans certains cas ont un rôle purement documentaire.

Dans ***WordNet***, la synonymie est contextuelle : « deux expressions sont synonymes dans un contexte linguistique C si la substitution de l'une par l'autre dans C ne modifie pas la valeur de vérité. Par exemple, le fait de substituer *plank* à *board* modifie rarement la valeur de vérité dans des contextes liés à la charpenterie, mais cette substitution serait totalement inappropriée dans d'autres contextes de *board*[33]. » (*ibid*., p. 6).

---

[29] WordNet est disponible par ftp anonyme depuis ftp.cogsci.princeton.edu ou ftp.ims.uni-stuttgart.de (sept. 1997). Il existe en différentes versions pour Unix, PC Windows et Macintosh.

[30] La terminologie de ***WordNet*** identifie le sens d'un mot au concept sous-jacent.

[31] Ceci s'oppose aux approches constructivistes qui tendent à définir un sens en le décomposant en primitives de significations.

[32] ***WordNet*** n'existant pas à ce jour pour le français, tous les exemples sont empruntés à l'anglais. Les différents sens de *credit* distingués ici se retrouvent approximativement pour le nom français *crédit* : argent mis à disposition d'autrui (1), mérite (2), paiement différé (3).

[33] *Plank* et *board* sont synonymes dans le sens de *grosse planche*, mais *board* admet beaucoup d'autres sens : *tableau*, *cartonnage*, *comité*… (NDA).

### 4.1.2 Mettre les « sens » en réseau

Si le *synset* (ensemble de synonymes, dans la terminologie de ***WordNet***) sert d'identifiant pour un sens, la liste des mots qui le composent ne donne qu'une vue très partielle du concept sous-jacent. Les liens que ce synset entretient avec d'autres synsets la complètent.

***WordNet*** est conçu comme un réseau lexical. Les synsets en sont les nœuds. Ils sont reliés entre eux par des relations d'hyponymie, d'antonymie, de méronymie[34], d'implication ou de dérivation morphologique[35]. La figure ci-dessous montre de manière simplifiée[36] comment le premier sens de *credit* (*crédit*) se situe par rapport aux synsets voisins : c'est un hyponyme de *asset* (*avoir*), un hyperonyme lointain[37] de *credit-card* (*carte de crédit*), un antonyme de *cash* (*argent comptant*).

Les relations qui structurent ***WordNet*** n'ont pas toutes le même statut. La synonymie joue un rôle central dans la mesure où elle est interne aux nœuds et constitutive des synsets. Elle s'oppose à toutes les autres relations, qui relient les mots les uns aux autres. Cela revient à distinguer deux niveaux de relations : les relations *lexicales*[38] qui relient respectivement entre eux les mots et les relations *sémantiques* qui relient entre eux les sens de mots, c'est-à-dire les synsets ou concepts.

Par ailleurs, les relations d'hyponymie et de méronymie se distinguent des autres parce qu'elles construisent une hiérarchie entre les nœuds qu'elles relient. Ces liens hiérarchiques déterminent des possibilités d'héritage au sens où les nœuds héritent certaines propriétés des nœuds qui les dominent. Dans l'exemple ci-dessus, si le nœud COIN porte une propriété héritable (le fait d'être composé de métal, par exemple, qui pourrait être représenté par un lien méronymique de matière entre les nœuds METAL et COIN), les nœuds NICKEL et DIME, héritent cette propriété de leur hyperonyme.

### 4.1.3 Quelques chiffres

La taille du vocabulaire couvert suffit à donner la mesure de l'ambition qui a présidé à la construction de ce réseau. ***WordNet*** comporte[39] 95 600

---

[34] Relation de partie à tout. (Cf. section 4.2.1.).

[35] Nous mettons l'accent sur les aspects sémantiques et nous ne considérons pas ici les liens de morphologie flexionnelle.

[36] N'est reproduite ici qu'une portion du sous-réseau concerné. Pour ne pas surcharger la figure, un synset est représenté par un mot clé, emprunté à la liste des mots qui le définit et noté en petites majuscules.

[37] La chaîne d'hyponymies complète est la suivante (les synsets et la relation d'hyponymie sont respectivement notés entre accolades et par le signe « < ») : {*credit card, charge card, charge plate, bank card*} < {*open-end credit, revolving credit, charge account credit*} < {*consumer credit*} > {*credit line, line of credit, bank line, line, personal credit, personal line of credit*} > {*credit*}.

[38] Nous reprenons ici la terminologie de ***WordNet***.

[39] Les chiffres que nous citons sont ceux que donnent (Miller *et al.,* 1993). Ce sont des

unités lexicales différentes : 51 500 mots simples et 44 100 expressions (*collocations*). À ces mots sont associés quelques 70 100 sens différents. Le tableau 3.1 montre comment ces unités et sens se répartissent.

*Tableau 3.1*

| | **Noms** | **Verbes** | **Adjectifs** |
|---|---|---|---|
| **Nombre d'unités lexicales** | 57 000 | 21 000 | 19 500 |
| **Nombre de sens** | 48 800 | 8 400 | 10 000 |
| **Nombre de catégories générales** | 25 | 14 | |

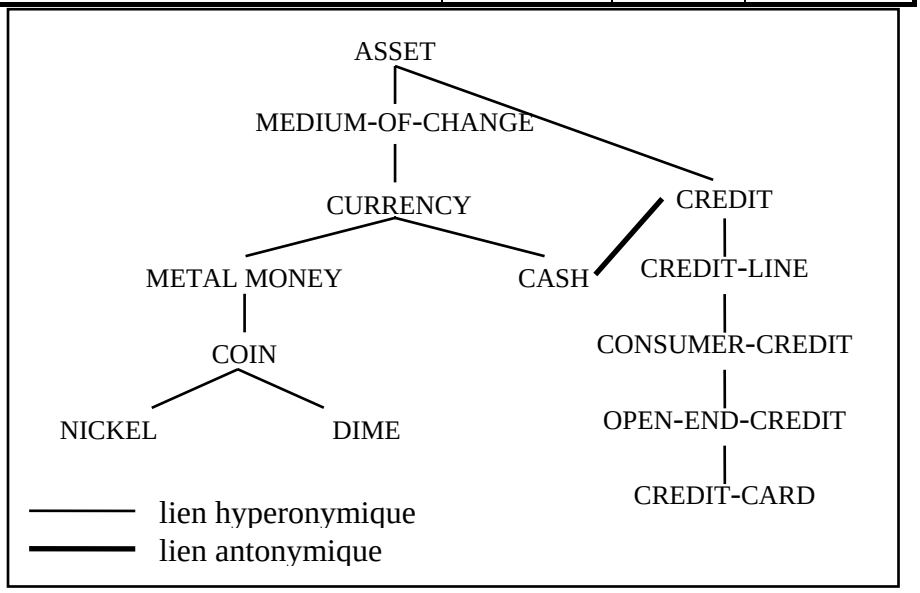


*Figure 3.4.— Exemple de sous-hiérarchie de **WordNet**.*

## *4.2 Une structure riche et différenciée*

***WordNet*** décompose le lexique en cinq catégories : noms, verbes, adjectives, adverbes et mots fonctionnels[40]. Chacune de ces catégories a sa propre structure interne. « Ce sont des expériences sur les associations de mots qui ont mis en évidence à l'origine que l'organisation […] varie d'une catégorie syntaxique à l'autre. » (*ibid*.).

---

approximations, ce qui explique l'inexactitude des totaux. ***WordNet*** continue de croître.

[40] Cette dernière catégorie n'est toutefois pas intégrée à ***WordNet*** (NDA).

### 4.2.1 Des hiérarchies de noms

L'ensemble des noms, qui comporte des formes simples et des mots composés mais pas de noms propres, est organisé autour de la relation d'hyponymie qui se définit comme suit : « on dit qu'un concept représenté par le synset {x, x',…} est l'hyponyme du concept représenté par le synset {y, y',…} si les locuteurs dont l'anglais est la langue maternelle acceptent les phrases du type *Un x est une sorte de y*. » (ibid., p. 8) Miller (1993, p. 17) donne un exemple de chaîne hyponymique :

> *televangelist* < *evangelist*< *preacher* < *clergyman* < *spiritual leader* < *person*[41]

La structure induite est en fait un ensemble de 25 hiérarchies dominées par des catégories sémantiques générales (*unique beginner*) : *person* dans l'exemple ci-desus ; *possession*, hyperonyme direct de *asset*, pour la sous-hiérarchie représentée par la figure 3.1. Au sein d'une hiérarchie, la hauteur est variable : selon les zones du lexique concernées, les synsets les plus bas se situent à 3, à 10, parfois même à 12 niveaux d'écart du sommet. De fait, si le vocabulaire technique se prête souvent bien à ce type d'organisation[42], il est plus difficile de définir des chaînes hyponymiques entre les mots de la langue courante (Kleiber et Tamba, 1990) : dans l'exemple ci-dessus, on peut se demander si tous les prédicateurs (*preacher*) sont effectivement des ecclésiastiques (*clergyman*).

Il faut souligner que les liens hyponymiques d'une taxonomie lexicale ne représentent pas une distance uniforme. Dans la pratique, on peut donc distinguer des grandes catégories générales qui forment le sommet des différentes hiérarchies ou la totalité des synsets. Il est difficile d'établir des distinctions intermédiaires. G. Miller (1993, p. 17) considère qu'il existe un niveau fondamental (*basic level*) qui permettrait de définir des catégories génériques ou fondamentales : situé quelque part entre le sommet et la base de la hiérarchie, c'est le niveau qui est le plus riche en relations. Dans la pratique ce niveau fondamental n'est pas clairement identifiable.

Cette structure hiérarchique peut être parcourue de haut en bas ou de bas en haut. À partir d'un sens donné, on peut ainsi retrouver ses ancêtres (hyperonymes directs et indirects), ses descendants (hyponymes directs ou indirects) mais aussi ses frères (coordinates).

Outre leur place dans cette structure hiérarchique, les sens des noms se définissent par des propriétés : leurs attributs, leur composition et leurs fonctions. La composition est décrite par différents types de relations méronymiques dans ***WordNet*** : les relations de composant à objet

---

[41] Dans x < y, le mot *x* est donné comme l'hyponyme du mot *y*. On aurait pour le français la séquence suivante : *télé-évangéliste* < *évangéliste* < *prédicateur* < *ecclésiastique* < *chef spirituel* < *personne*.

[42] C'est particulièrement vrai de la botanique ou de la zoologie, domaines où la connaissance est traditionnellement organisée selon les catégories de l'espèce, du genre, du taxon…

composé (*branche* / *arbre*), d'élément à ensemble (*arbre* / *forêt*) et de matière (*arbre* / *bois*). En revanche, les attributs (un arbre peut être grand, vieux…) et les fonctions (*une hache sert à couper…*) ne sont pas représentés dans ***WordNet***. Ce sont en effet des relations trans-catégorielles qui devraient à terme relier les hiérarchies de noms aux réseaux des adjectifs ou des verbes.

### 4.2.2 Des classes d'adjectifs

Les synsets d'adjectifs comprennent essentiellement des adjectifs qualificatifs[43], même si des noms ou locutions prépositionnelles utilisées comme modifieurs y figurent également. Ces adjectifs ne s'organisent pas comme les noms. Pour les adjectifs, il n'existe pas de relation hiérarchique comme l'hyponymie.

La relation fondamentale structurant l'espace des adjectifs est l'antonymie. Cette relation symétrique, mise en évidence par des tests psycholinguistiques sur les associations de mots, est difficile à formaliser. Les auteurs retiennent l'idée que les adjectifs antonymes expriment deux valeurs opposées d'un même attribut.

Partant cependant du constat que certains adjectifs proches par le sens (*heavy* et *weighty*[44], par exemple) ont des antonymes différents (*light* et *weigthless*[45]) et que beaucoup d'adjectifs qualificatifs (*ponderous*[46]) n'ont pas d'antonymes directs, la structure retenue est celle de classes d'adjectifs similaires entre eux, ces classes étant organisées autour d'adjectifs pôles qui peuvent s'opposer à d'autres pôles par des liens d'antonymie. *heavy* et *light* sont donc considérés comme antonymes, mais *ponderous*, qui est similaire à *heavy* et qui n'a pas d'antonyme direct n'est qu'un antonyme indirect de *light*.

### 4.2.3 Des réseaux de verbes

Comme les noms et les adjectifs, les verbes sont regroupés en synsets. Ceux-ci comportent des formes simples mais aussi des tournures verbales, comme *look up*, qui sont très fréquentes en anglais. Les synsets se répartissent eux-mêmes en 15 catégories générales (14 pour les actions et événements ; 1 pour les états).

La relation centrale pour le réseau des verbes n'est ni l'hyponymie, ni l'antonymie, mais l'implication. ***WordNet*** en distingue quatre types : la

---

[43] ***WordNet*** distingue les adjectifs qualificatifs des adjectifs relationnels. On a vu au chapitre 1, l'intérêt de ce types de distinction pour le traitement de ***Enfants***. Les adjectifs relationnels sont considérés comme des « variantes stylistiques » de noms : ils se définissent par rapport à ces noms auxquels ils sont liés. Nous mettons ici l'accent sur les seuls adjectifs qualificatifs.

[44] *lourd* et *pesant*, respectivement.

[45] *léger* et *de peu de poids*.

[46] *massif, pesant.*

cause (*give* / *have* : *donner* / *avoir*), la présupposition (*succeed* / *try* : *réussir* / *essayer* ou *untie* / *tie* : *dénouer* / *nouer*), l'inclusion (*snore* / *sleep* : *ronfler* / *dormir* ou *buy* / *pay* : *acheter* / *payer*) et la troponymie[47] (*limp* / *walk*, *boiter* / *marcher*).

Soulignant toutefois la complexité de la sémantique des verbes et la difficulté de définir une sémantique proprement différentielle, les auteurs de WordNet reconnaissent la moindre maturité du réseau des verbes. Dans la pratique, les travaux qui exploitent ce réseau des verbes à des fins de désambiguïsation lexicale s'en tiennent souvent aux grandes catégories sémantiques (Basili *et al*., 1997).

## 5. TABLER SUR L'EXISTANT

Les ressources lexicales existantes ont chacune leurs faiblesses. Dès lors qu'elles visent une couverture un peu large du lexique, elles reposent sur des approximations. Dans ***WordNet***, les sens représentés par les synsets sont souvent difficiles à maîtriser pour qui n'est pas lexicographe professionnel et ils comportent une part importante d'arbitraire. C'est le cas pour tous les dictionnaires. Les catégories sémantiques très générales, à l'inverse, sont souvent peu contestables car peu discriminantes. La hiérarchie des noms, la partie la plus stable du réseau, repose sur des chaînes d'hyponymie qui pour la langue générale sont le plus souvent approximatives. La structuration des réseaux des adjectifs ou des verbes paraît moins solide.

Pourtant, l'apparition de ressources lexicales de taille importante, aussi imparfaites soient-elles, a donné le coup d'envoi des travaux de sémantique à partir de corpus. Ce sont des dictionnaires sur support informatique ou des thesaurus électroniques comme ***WordNet*** qui ont permis de mettre au point de nouvelles méthodes de désambiguïsation automatique (cf. IV-3). Et c'est l'utilisation même de ces ressources qui permettra d'en améliorer la conception. La lexicographie électronique à proprement parler n'en est encore qu'à ses débuts : de nouveaux moyens de stockage et d'investigations induisent de nouvelles structures et organisations de données, lesquelles donnent à voir de nouveaux phénomènes.

Ceci nous amène à souligner avec inquiétude l'absence de ressources similaires pour le français[48]. Si la recherche sur les corpus en français peut sans doute tirer profit de l'expérience anglo-saxonne pour éviter certains tâtonnements, des problèmes spécifiques se posent pour chaque langue, qui imposent certains ajustements, voire la mise au point de

[47] Un verbe x est un troponyme d'un verbe y si on peut dire que *x, c'est y d'une certaine manière.*

[48] ***EuroWordnet***, un projet de construction d'un ***WordNet*** multilingue a été lancé en mars 1996 (Vossen, 1996). Il concerne initialement l'allemand, l'italien et l'espagnol. La France accuse un certain retard.

méthodes particulières ou le développement d'outils spécifiques. L'absence de ressources lexicales informatisée pour le français est déjà un frein pour tous les traitements sémantiques. Faute de moyens, la plupart des travaux français s'intéressent à l'acquisition de connaissances à partir de corpus (cf. chapitre VIII, section 5).