CHAPITRE IV

DES MOTS AUX SENS : SEMANTIQUE EN CORPUS

1. DEFINITIONS ET ENJEUX

Les travaux sur corpus dans le domaine sémantique foisonnent. D'une expérience à l'autre, l'objectif est toujours d'accéder au sens que véhicule le corpus mais ces travaux, pour la plupart assez ponctuels, ont des visées extrêmement variées et s'appuient sur des méthodes fort différentes. Le présent chapitre cherche à faire apparaître à la fois l'unité et les contrastes d'un domaine aujourd'hui très productif. Les travaux s'inscrivent en fait dans des perspectives très différentes : nous en dressons une typologie schématique ci-après. Nous décrivons ensuite deux exemples d'applications représentatives des travaux de sémantique sur corpus. En 2, nous nous appuyons sur les travaux de G. Grefenstette pour montrer le parti que la lexicographie spécialisée peut tirer de l'exploitation systématique de corpus enrichis. La partie 3, plus prospective, met l'accent sur la recherche documentaire et sur l'apport des techniques de désambiguïsation lexicale dans ce contexte. Nous terminons, en 4, en montrant que ces deux expériences, qui s'opposent par leurs méthodes, relèvent en fait d'une même démarche empirique.

1.1 Un objectif commun : accéder au sens

Des corpus porteurs d'annotations sémantiques commencent à voir le jour, mais on n'en est cependant qu'aux balbutiements, que ce soit pour la

constitution de ces corpus ou pour leur exploitation.

Pourtant — cela transparaît dans les exemples des chapitres I et II— les préoccupations sémantiques occupent une place importante dans l'exploitation des corpus, que l'on cherche à identifier la terminologie d'un domaine technique, à traduire des expressions figées, à repérer les thèmes abordés par différentes catégories de répondants à une enquête d'opinion, le genre des textes, etc. Si de nombreuses études portent sur la facture même des corpus et la langue employée, le texte demeure un message porteur d'information et l'on ne cesse d'interroger les corpus sur le sens qu'ils véhiculent.

Le présent chapitre met l'accent sur l'exploitation sémantique des corpus, laquelle peut porter aussi bien sur des corpus nus que sur des corpus étiquetés et arborés. Sur les deux exemples de l'aide à la lexicographie et de la recherche d'information, il tente de montrer dans quelle mesure et à quelles fins on peut accéder au sens véhiculé par les phrases ou les textes d'un corpus.

1.2 Des applications variées

L'analyse sémantique intéresse des domaines et des publics extrêmement divers. On peut identifier trois principaux types d'applications : l'analyse de contenu, l'acquisition de connaissances et la recherche documentaire.

1.2.1 Analyse de contenu

L'analyse sémantique vise tout d'abord à rendre compte du « contenu » des corpus, s'inscrivant en cela dans une longue tradition à la fois littéraire, stylistique, historique et sociologique. Que l'objectif soit de rendre compte des propriétés esthétiques, de retracer une évolution historique ou de décrire un moment de l'histoire, de caractériser les discours de certaines catégories de population, il s'agit d'explorer le contenu des corpus en tant que tel pour en repérer à la fois les thèmes dominants et leur agencement.

Les études thématiques s'intéressent principalement au lexique. On a ainsi montré comment évolue dans *À la recherche du temps perdu* le champ sémantique du temps, lequel devient de plus en plus présent et de plus en plus sombre au fur et à mesure que l'on avance dans l'œuvre (Brunet, 1983), comment se transforment les idées révolutionnaires dans le discours de Roselière, quelles sont les préoccupations que mettent principalement en avant les jeunes dans les enquêtes d'opinion (Lebart et Salem, 1994).

Au delà de la seule étude du vocabulaire, l'ambition de M. Pêcheux avec l'analyse du discours est de mettre en évidence, sous la diversité des formes rhétoriques de surface, les phrases élémentaires ou « de base » d'un discours. Il s'agit par exemple pour Pêcheux et ses collègues

de mettre en évidence l'ambiguïté idéologique du « rapport Mansholt » (Maingueneau 1991).

Le recours aux méthodes statistiques a déjà permis de renouveler les études thématiques (Brunet, 1991), mais l'existence de corpus étiquetés et surtout arborés ouvre de nouvelles perspectives en matière d'analyse de contenu.

1.2.2 Recherche documentaire

Dans le prolongement des analyses thématiques, l'analyse sémantique de corpus intéresse également la recherche documentaire. Les codifications traditionnelles des bibliothécaires reflètent les thèmes principaux des ouvrages. Avec l'essor des besoins en traitement de l'information et le développement d'une véritable industrie, on cherche aujourd'hui à développer des outils automatiques.

Quels que soient les textes — ouvrages, parties d'ouvrages, articles ou même dépêches, écrits dans une ou plusieurs langues, documents techniques ou non —, quand on a affaire à un nombre important de textes, il faut faire du tri. Deux voies sont possibles. Les document peuvent être classés *a priori* en groupes homogènes, le plus souvent thématiques, mais le tri peut aussi se faire *a posteriori*, en fonction d'un objectif spécifique, par l'extraction ciblée d'un sous-ensemble de textes pertinents au regard de cet objectif. La première direction soulève deux difficultés. Si l'éventail des catégories est donné au préalable, il faut identifier les indices permettant d'associer à un texte une ou plusieurs catégories (on parle alors de catégorisation de textes). Mais si le jeu de catégories n'est pas donné, il faut également déterminer les critères de classement (classification de textes). Dans la seconde direction, le critère de choix est fixé par l'utilisateur qui formule une requête (les textes portant sur l'aéronautique, par exemple), mais il faut repérer les multiples formes sous lesquelles ce thème peut être exprimé dans la base de textes interrogée.

Les premiers outils de recherche documentaire reposaient et reposent encore souvent sur des mots clefs censés refléter le « contenu » du document. Toute la question est alors de déterminer quels sont les mots les plus représentatifs d'un document et de guider l'utilisateur dans la formulation de sa requête si les mots clefs qu'il donne comme critère de recherche sont trop ou trop peu spécifiques. Les travaux en analyse sémantique de corpus permettent aujourd'hui d'envisager de réelles améliorations dans le domaine de la recherche documentaire (voirsection 3).

1.2.3 Acquisition de connaissances

L'analyse sémantique de corpus vise enfin à acquérir des connaissances à partir de corpus. Partant du constat que, dans nos sociétés modernes,

l'écrit est le principal véhicule de l'information et des connaissances et que, hors des domaines formels pour lesquels ont été conçus des langages formels, mathématiques ou logiques, ces connaissances sont toujours exprimées en langage naturel, on cherche à développer des méthodes pour extraire et donc acquérir les connaissances des corpus. Il s'agit ni plus ni moins de proposer des techniques de « lecture » rapide et automatique des corpus.

Les connaissances ainsi extraites servent souvent à construire les bases de connaissances lexicographiques que sont les dictionnaires, thesaurus et terminologies, qu'elles soient de langue générale ou spécialisées, monolingues ou bilingues. Nous développons cet aspect ci-dessous (section 2). Il s'agit également de modéliser l'ensemble des connaissances constituant un domaine spécialisé. Un corpus portant sur l'aéronautique doit ainsi permettre d'identifier les différentes pièces composant un avion et leurs agencements, leur usage habituel, les dysfonctionnements susceptibles de se produire, etc. Le modèle de connaissances ainsi construit donne alors une vue schématisée du domaine. Celle-ci est précieuse pour le développement d'applications évoluées comme les outils de diagnostic de panne, des outils de visualisation, des simulateurs de vols, des systèmes d'aide au pilotage, etc. De la même manière, (Bouaud*.et al.*, 1997) exploite ***Menelas*** pour aider à la construction de l'ontologie du domaine des maladies coronariennes. L'extraction des informations véhiculées par un corpus sert encore à alimenter des bases de données. L'exploitation d'un corpus de dépêches portant sur le terrorisme permet ainsi de stocker les données relatives aux événements terroristes dans (Appelt *et al.*, 1993).

Ce panorama, nécessairement schématique, montre que l'analyse sémantique aborde les corpus tour à tour comme un objet à décrire (analyse de contenu), comme un ensemble de documents à classer et à retrouver (recherche documentaire) ou comme une source de connaissances (acquisition de connaissances). La diversité des applications visées montre également que, pas plus qu'en matière d'étiquetage ou de structuration de corpus, il n'existe de consensus en matière sémantique lorsqu'il s'agit de rendre compte du « sens ». Le sens de la recherche documentaire (ensemble de thèmes) ne correspond pas au sens que l'analyse du discours cherche à exhiber sous la forme de phrases de base et pas davantage au sens des mots et locutions que les lexicographes tentent de décrire. Nous développons ci-dessous en 2 et 3 deux exemples d'applications qui s'inscrivent respectivement dans le champ de l'acquisition de connaissances — en l'occurrence, lexicographiques — à partir de corpus spécialisés et dans celui de la recherche documentaire. Par leur démarche empirique (nous y revenons en 4), ces exemples nous paraissent représentatifs des travaux actuels en matière d'exploitation sémantique de corpus.

2. CONSTRUIRE AUTOMATIQUEMENT DES ENTREES DE DICTIONNAIRE

Le travail du lexicographe, pour la langue générale, consiste le plus souvent à fusionner et mettre à jour des sources antérieures existantes. Mais élaborer des dictionnaires pour une langue spécialisée suppose de cerner la langue considérée. Le lexicographe doit généralement se familiariser avec le domaine par la lecture des textes produits par les acteurs du domaine, puis compléter ses connaissances par des entretiens avec les experts du domaine. Le coût de ce travail est indubitablement un frein à l'élaboration de ces dictionnaires spécialisés et la perspective de pouvoir les construire automatiquement ou semi-automatiquement à partir de corpus est alléchante. L'hypothèse sous-jacente est qu'il est possible d'inférer une description de la langue considérée à partir des observations faites sur le corpus.

Pour G. Grefenstette, cette perspective est réaliste (1994a, p. 135) : les sens généraux des mots peuvent être identifiés à partir des schémas syntaxiques et lexicaux dans lesquels ils figurent en corpus et nous avons les moyens de repérer objectivement ces sens et de les décrire ». Ses travaux montrent qu'il est possible de construire automatiquement des ébauches d'entrées de thesaurus qui peuvent aussi bien servir de base à un lexicographe pour la rédaction d'entrées de dictionnaires.

Nous présentons dans un premier temps les résultats qu'il obtient. Nous en soulignons l'intérêt lexicographique. Nous décrivons ensuite les méthodes qui permettent d'obtenir ces résultats automatiquement à partir de corpus. Nous terminons en indiquant les limites de cette approche.

2.1 Des ébauches d'entrées de dictionnaires

Nous présentons ci-dessous les exemples d'entrées de dictionnaire que donne G. Grefenstette (*ibid.*, annexe 5) pour les mot *growth* (*croissance*), *therapy* (*thérapie*) et *year* (*année*). Elles suivent le schéma suivant :

> **<Nom vedette>** :: [<données quantitatives>] <NOM DU cORPUS D'ORIGINE> ***Relat***.[1] <liste des noms voisins>. ***Vbs***. [2] <liste des verbes opérateurs>. ***Exp***. [3] <liste des expressions et de leurs expressions voisines>. ***Fam***. [4] <liste des variantes >.

Ces entrées ont été construites entièrement automatiquement à partir de deux corpus spécialisés différents (MED ou MERGERS, cf. *infra*).

> **Growth** :: [284 contexts, frequency rank : 25] MED ***Relat***. tumor **;** effect,

[1] Pour *related words.*
[2] Pour *verbs.*
[3] Pour *expressions.*
[4] Pour *family*.

tissue **;** antigen, protein, development. ***Vbs***. retard, stimulate, show, follow, enhance, accelerate. ***Exp***. growth hormone (cf. bone marrow, parathyroid hormone), growth rate (cf. growth retardation, folic acid), tumor growth (cf. body growth, tenuazonic acid), growth retardation (cf. dna content, body weight), body growth (cf. tumor growth, body weight).

Therapy :: [256 contexts, frequency rank 28] MED ***Relat***. test **;** response, treatment **;** procedure, operation, drug, chemotherapy, dose, administration. ***Vbs***. use, respond, follow, remain, receive, combine. ***Exp***. radiation therapy (cf. survival rate, cancer chemotherapy), steroid therapy (cf. inclusion disease, cancer chemotherapy), hormone therapy (cf. intra-arterial infusion, steroid therapy), corticosteroid therapy (cf. connective tissue, plama concentration). ***Fam***. therapeutic.

Year :: [103 contexts, frequency rank 93] MED ***Relat***. woman **;** child, patient, day **;** week, month, hour. ***Vbs***. age, occur, follow. ***Exp***. year period (cf. survival rate, hormone therapy).

Growth :: [320 contexts, frequency rank : 139] MERGERS ***Relat***. level, increase, gain **;** loss **;** performance, return, rise, decline, flow, expansion. ***Vbs***. say, expect, slow, accelerate, maintain, sustain, forecast, continue. ***Exp***. rapid growth (cf. buy-out bid, raise capital), profit growth (cf. electronics group, total revenue), growth rate (cf. profit margin, future performance), growth potential (cf. company spokeswoman, board seat), future growth (cf. speciality chain, bottom line).

Ces entrées ne ressemblent guère à des entrées habituelles de dictionnaire[5]. Pourtant, elles constituent un ensemble d'indications qui peut guider le lexicographe dans son travail de rédaction. Elles comportent six rubriques, les quatre dernières étant optionnelles.

2.1.1 Des données quantitatives

Le nombre de contextes ou d'occurrences du nom vedette et son rang dans l'ordre de fréquences décroissantes renseignent sur son poids dans le corpus. Les noms les plus fréquents du corpus médical (par ordre décroissant *cell*, *patient*, *effect*, *study*, *case*) sont en effet représentatifs du domaine considéré. Sur l'exemple ci-dessus, on constate que *growth* et *therapy* sont ainsi nettement plus fréquents que *year*. De surcroît, on sait que le rang des noms d'un corpus donne une indication sur le degré de spécificité ou de généralité de ces noms (Srinivasan, 1992). Le fait que *patient* soit plus fréquent que *child* ou *woman* ; *treatment* plus fréquent que *therapy*, lui-même plus fréquent que *chemotherapy* paraît en effet suggérer que *patient* fonctionne dans le corpus médical comme

[5] Elles s'apparentent davantage, comme le souligne G. Grefenstette, à des entrées de thesaurus.

l'hyperonyme de *child* ou *woman* ou que la chimiothérapie est une sorte de thérapie et de traitement.

2.1.2 Le corpus d'origine

Cette indication (ici MED ou MERGERS) est évidemment importante dans la mesure où il s'agit de décrire des langues spécialisées à partir de corpus. Les trois premières entrées sont construites à partir d'un corpus de résumés médicaux (MED). La dernière, à partir d'un ensemble d'articles du *Wall Street Journal* portant sur la fusion d'entreprises (MERGERS). Le contraste entre les deux entrées de *growth* montre deux sens spécialisés différents.

2.1.3 Les noms voisins

Cette liste, qui est introduite par le mot clef *Relat.*, comporte des noms donnés comme sémantiquement proches du nom vedette. Dans le corpus financier, *growth* se trouve au voisinage d'une dizaine de noms : *level, increase, gain **;** loss **;** performance, return, rise, decline, flow, expansion*. *S*oulignons la cohérence de cette liste[6]. Elle comporte essentiellement des synonymes ou des pseudo-synonymes (*increase, gain, rise, expansion*) et quelques antonymes (*loss, decline*). Même si le lien de *growth* avec *level*, *performance* et *flow* est moins évident, le rapprochement de ces termes paraît néanmoins assez judicieux. Seul *return* surprend. La liste des voisins est structurée en trois parties séparées par des points virgules. Sont ainsi distingués les voisins qui sont plus fréquents, aussi fréquents et moins fréquents que le mot vedette, cette indication pouvant refléter le degré de généralité. Pour le lexicographe, cette liste donne un premier aperçu des relations lexicales autour du nom vedette, relations dont il n'est pas évident de se faire une idée *a priori*, à la lecture du corpus ou même à partir de concordances. Cette liste doit être contrôlée, parfois émondée ou complétée : la liste des voisins de *year* semble peu satisfaisante, par exemple. Le retour aux contextes permet de vérifier le sens dans lequel les mots sont employés. Dans tous les cas, cette liste demande à être interprétée pour que soit identifiée la nature des relations lexicales sous-jacentes.

2.1.4 Les verbes opérateurs

Ces verbes sont introduits par le mot clef *Vbs*. Il s'agit des verbes auxquels le nom vedette est régulièrement associé, comme sujet, objet direct ou complément prépositionnel. Les verbes sont classés par ordre de fréquence décroissante. Cette rubrique renseigne sur les emplois du

[6] Le principe du calcul des similarités qui permet de construire cette liste est exposé au chapitre VIII.

nom vedette et les relations dans lesquelles il entre. On constate ainsi que la croissance (*growth*) dans le corpus financier, est quelque chose dont le rythme évolue (*slow, accelerate, maintain, sustain, continue*), mais aussi quelque chose qui se prévoit (*expect*, *forecast*). En termes de fréquences, c'est surtout quelque chose dont on parle ou qui donne des informations (*say*)[7]. En fait, cette rubrique des verbes opérateurs donne une première indication synthétique des contextes d'emplois du nom vedette. Le fait que *age* (*âgé de*) figure parmi les verbes associés à *year* explique la présence surprenante à première vue des noms de personnes (*women*, *child*, *patient*, etc.) aux côtés des termes de durée (*day*, *week*, *month*, etc.). C'est, semble-t-il, l'importance des contextes du type *woman aged of thirty years* qui rapproche *woman* et *year*.

2.1.5 Les expressions

La liste des expressions nominales les plus fréquentes dans lesquelles entre le mot vedette donne une autre indication contextuelle. Comme la précédente, cette rubrique (introduite par *Exp.*) permet par exemple de contraster les emplois de *growth* dans la langue médicale et dans la presse financière. Dans les deux cas, on parle du rythme de la croissance (*growth rate*, *growth retardation*, *rapid growth*), mais l'objet de la croissance diffère (*tumor, body* dans un cas, *profit* dans l'autre). À chaque expression sont associées une ou plusieurs expressions voisines à titre de documentation. G. Grefenstette souligne ainsi l'écart d'emploi d'une expression commune aux deux corpus (*growth rate*) : dans un cas, *growth rate* est associé à *growth retardation* tandis que dans l'autre corpus, le taux de croissance est associé à des considérations de profit et de performance.

2.1.6 Les variantes

Cette dernière rubrique (introduite par le mot clef *Fam*.), souvent absente, donne des variantes morphologiques du nom vedette, généralement un équivalent adjectival ou verbal (*therapy/therapeutic*, *bile/biliary*, *excretion/excrete*, *reduction/reduce*). Il est souvent précieux pour un non spécialiste du domaine de repérer quelles sont, dans l'ensemble des dérivations possibles en langue, celles qui sont attestées dans le corpus ou au contraire de constater qu'un équivalent possible ne semble pas employé. Ainsi l'entrée de *blood* (*sang*) ne mentionne-t-elle pas *bloody* (*sanglant*) qui, de fait, n'a guère un sens médical. On trouve également sous cette rubrique des variantes orthographiques (*adeaminase/a-deaminase*). Dans certains cas, cette rubrique regroupe non pas des variantes à proprement parler mais des mots qui appartiennent à la même famille dérivationnelle (*lymphocyte/lymph/lymph node/lymphatic/lymphoid*) sémantique.

[7] Pour savoir si *growth* figure comme sujet et/ou comme objet du verbe *say*, il faut revenir au corpus.

Des corpus porteurs de sens

Le recours aux corpus, plutôt qu'à l'introspection, est chose ancienne pour la lexicographie spécialisée et il est clair que les entrées ainsi constituées automatiquement demandent à être retravaillées par un lexicographe. Le travail de G. Grefenstette montre cependant toutes les possibilités que le traitements automatiques de corpus ouvrent désormais. Rappelons en effet que les entrées données ci-dessus ont été engendrées de manière entièrement automatique. Ces entrées constituent des ébauches ou un premier dégrossissage qui donnent au lexicographe une vue synthétique sur le poids (données quantitatives) et le fonctionnement syntagmatique (expressions et verbes opérateurs) ou paradigmatique (voisins et variantes) d'un mot dans le corpus considéré.

2.2 Une méthode entièrement automatique

Ces entrées ne sont pourtant pas de qualité égale. L'entrée de *year* paraît plus difficile à exploiter que celles de *growth*. En règle générale, on constate que plus les noms sont techniques et fréquents, meilleure est leur description. Pour apprécier la pertinence des informations extraites et savoir interpréter des résultats parfois surprenants, il importe de comprendre par quelles méthodes et dans quelles conditions ces entrées ont pu être construites à partir des corpus.

2.2.1 Une seule donnée, le corpus

En matière de données, la méthode repose sur le corpus et sur le corpus seulement. Dans la mesure où il est exploité comme source de connaissances pour décrire une langue spécialisée, il est primordial de partir d'un corpus homogène et représentatif de cette langue (voir chapitre VII), mais en tant que telle la méthode d'extraction de G. Grefenstette est indépendante du domaine traité. Au-delà des corpus médicaux et financiers cités ci-dessus, cette méthode a été testée « avec succès » sur « plus de 20 corpus de 1 à 6 millions de caractères » (Grefenstette, 1993), soit approximativement de 150 000 à 850 000 mots. Ces corpus sont préalablement étiquetés.

La construction de ces entrées de dictionnaire ne fait appel à aucune connaissance sémantique. C'est là le point fort de la méthode qui repose sur des techniques de bas niveau (*knowledge-poor technique*s) en ce sens que le processus d'extraction repose entièrement sur des traitements morpho-syntaxiques et statistiques du corpus[8].

[8] « Nous parlons de traitement de bas niveau parce que c'est une approche des textes qui ne nécessite pas qu'une modélisation sémantique des connaissances du domaine soit préalablement construite à la main » (Grefenstette, 1994a, p. 3).

2.2.2 Un ensemble de traitements simples

Le traitement du corpus est effectué par le logiciel SEXTANT (Grefenstette, 1994a) qui traduit dans un premier temps le corpus préalablement étiqueté en un ensemble de relations de dépendances syntaxiques. L'accent est mis sur les noms et ne sont conservées que les relations entre un nom d'une part et un adjectif, un verbe ou un autre nom, d'autre part. En simulant ce traitement sur les extraits de ***Menelas*** donnés ci-dessous, on obtient comme contextes pour le nom *épisode* ses relations avec les mots suivants[9] : *présenter* (OBJ), *survenir* (SUJ), *douloureux*, *précordial*, *hyperthermique*, *effort*, *repos*.

> Traité médicalement, il a déjà *présenté* à plusieurs reprises des ***épisodes*** *douloureux précordiaux d'effort* et de *repos*.

> Depuis cette époque on ne note aucune récidive d'angor jusqu'il y a 8 jours où il a *présenté* un ***épisode*** de *précordialgie survenant* à l'effort, durant environ 45 minutes, sans irradiation[10].

> On notait par ailleurs la *survenue* d'un ***épisode*** *hyperthermique*, probablement en rapport avec une mise en place prolongée d'une voie veineuse.

Le nombre de contextes d'un nom est donc le nombre de relations de dépendance dans lesquelles il entre. C'est sur la base d'un corpus vu comme un ensemble de contextes que sont calculées toutes les informations syntagmatiques et paradigmatiques étudiées plus haut.

Les relations syntagmatiques sont données par les contextes eux-mêmes : les rubriques des verbes opérateurs et des expressions regroupent respectivement les contextes verbaux et nominaux du nom vedette. Le logiciel se contente de trier les listes par ordre de fréquence et d'éliminer les contextes trop peu fréquents ou syntaxiquement ambigus.

Les relations paradigmatiques sont calculées en comparant la liste des contextes de deux entités. Dans le cas du voisinage des noms, l'intuition sous-jacente est que deux noms sont voisins s'ils figurent dans les mêmes contextes ou s'ils partagent beaucoup de contextes. Par exemple, à supposer qu'on obtienne pour *symptomatologie* et *crise*, les listes de contextes suivantes :

> *symptomatologie* : présenter (OBJ), associer (OBJ), survenir(SUJ), douloureux, précordial, atypique, effort, problème

[9] Nous considérons ici que les mots ont été préalablement lemmatisés. Les marqueurs OBJ et SUJ indiquent respectivement que le nom figure en position objet ou sujet du verbe. Dans les résultats de G. Grefenstette, la nature des relations entre noms ou entre un adjectif et un nom n'est pas explicitée (1994a, p. 42).

[10] Nous n'avons pas considéré ici que les groupes prépositionnels *durant 45 minutes* et *sans irradiation* devaient être rattachés à *épisode*. Pour l'anglais, G Grefenstette résout le problème du rattachement du groupe prépositionnel par des règles *ad hoc* (*ibid.*).

crise : présenter (OBJ), prolonger (OBJ), suivre (SUJ), douloureux

la comparaison des distributions tend à montrer que *épisode* est plus similaire de *symptomatologie* que de *crise*. Formellement, les contextes d'un nom constituent un ensemble de propriétés (ses attributs) et le logiciel mesure le degré de similarité[11] entre deux noms sur la base du nombre d'attributs qu'ils partagent[12]. Dans la liste des voisins d'un nom vedette, on retient les noms qui en sont le plus similaires, à condition que, de manière réciproque, le nom vedette figure également en bonne position dans la liste des similaires de ceux-ci.

C'est sur le même principe que G. Grefenstette rapproche certaines expressions. Les expressions *radiation therapy* et *cancer chemiotherapy* sont associées parce qu'elles partagent un nombre de contextes qui est significatif étant donné le nombre total de contextes dans lesquels elles figurent. Pour ce calcul toutefois, G. Grefenstette ne retient pas les relations de dépendance binaire comme contexte, mais il prend un contexte plus large, la phrase.

C'est encore sur le même principe que sont calculées les variantes morphologiques. Le fait est que dans un paragraphe ou un document portant sur un sujet donné, une même notion s'exprime sous des formes diverses. Dans un document, on trouvera par exemple le verbe *réduire* et quelques lignes plus loin, la même idée reprise sous forme nominale (*réduction*). SEXTANT calcule donc des similarités entre les mots de sens plein du corpus en prenant comme contexte les numéros de documents dans lesquels ils figurent, puis il sélectionne ceux qui paraissent, sur une base graphique, être des variantes morphologiques.

Le principe général de SEXTANT est donc simple : il repose essentiellement sur le calcul de similarités. Tout l'intérêt vient d'une définition appropriée des contextes. Définir les contextes sur une base syntaxique plutôt que graphique revient à les filtrer au préalable et réduit le bruit engendré (Habert *et al.*, 1996 ; Grefenstette, 1996). Faire varier la taille des contextes permet de faire ressortir différents types d'association. Ces entrées de dictionnaires résultent d'un long travail d'expérimentation et d'une exploitation judicieuse de techniques simples.

2.3 Les limites d'une approche empirique

Pour bien utiliser un outil comme SEXTANT dans une perspective lexicographique, il est également important d'en connaître les limites. L'approche décrite ci-dessus présente certaines faiblesses. La rubrique la

[11] Nous entendons par *similarité* la relation existant entre deux choses *similaires*, c'est-à-dire « à peu près de même nature, de même ordre » (*Petit Robert*, édition de 1973). Nous avons recours à cet anglicisme parce que le mot *similitude* n'a pas le même sens que l'anglais *similarity* (« relation unissant deux choses exactement semblables » *Petit Robert*, édition de 1973).

[12] On trouve dans la littérature (Saporta, 1990) beaucoup de mesures de distances pour ce type de comparaison. G. Grefenstette retient une forme pondérée de l'indice de Jaccard qui rapporte le nombre d'attributs partagés par deux éléments au nombre d'attributs possédés en propre par l'un ou l'autre (1994a, p. 48-49).

moins satisfaisante est incontestablement celle des variantes qui mêle notamment les variantes orthographiques et dérivationnelles. L'algorithme de recherche des variantes morphologiques privilégie les variations qui ne portent pas sur l'initiale du mot et associe des mots qui ont seulement le même préfixe (*antigen* est associé à *antibody* mais pas à *gene*)[13].

Plus fondamentalement, les résultats dépendent de la qualité de l'analyse syntaxique. G. Grefenstette (1993) donne l'exemple curieux de *human cell* et *year period* associés à l'expression *cancer cell*. La décomposition des groupes nominaux du type *3 year period* est mal reconnue. Comme le système ne repère pas que *3* quantifie le seul *year*, il décompose *3 year period* en *[3 [year [period]]]* au lieu de *[[3 [year]]) period]*. Il analyse donc *3 year period* et *3 human cells* de la même manière et crée un rapprochement artificiel entre les deux expressions. Les erreurs d'analyse bruitent les résultats. L'exemple cité est suffisamment surprenant pour attirer l'attention du lexicographe, mais certaines erreurs de rattachement peuvent créer des rapprochements indus et néanmoins plausibles qui peuvent passer inaperçus. La fiabilité de l'analyse syntaxique est donc essentielle pour ce type de traitement. C'est la raison pour laquelle SEXTANT ne prend encore en compte que les relations de dépendance binaire dans le calcul des contextes et non les syntagmes nominaux de taille supérieure pour lesquels les risques d'erreur sont multipliés.

Le point essentiel demeure les contraintes d'une approche lexicographique consistant à inférer des propriétés en langue à partir des observations faites sur corpus, c'est-à-dire de ce qui est attesté. Cette approche repose sur l'hypothèse que le corpus est un reflet intéressant de la manière dont les mots sont effectivement employés. Cela suppose que le corpus soit homogène ou, du moins, que sa variation interne soit négligeable en regard des phénomènes étudiés. C'est une hypothèse forte, nous y revenons au chapitre VII. Le corpus détermine par ailleurs la couverture lexicographique : seuls les mots et les sens attestés peuvent être décrits puisque de la non-attestation, on ne peut jamais conclure qu'un mot est étranger à une langue de spécialité. Les mots faiblement représentés dans le corpus sont également difficiles à décrire. Les techniques utilisées par SEXTANT supposent que les mots aient un nombre « raisonnable » d'occurrences. La description construite à partir des 103 occurrences de *year* est nettement moins exploitable que celles de *growth* ou *therapy* qui portent sur deux fois et demi plus d'occurrences dans le corpus médical. La qualité et la fiabilité des descriptions lexicographique baissent avec le nombre de contextes dans lequel figurent les entrées, *i.e*. avec la quantité d'information disponible. Or des mots peu fréquents peuvent être des termes du domaine et certains emplois rares sont importants à décrire parce qu'ils sont difficiles à comprendre intuitivement.

On touche là aux limites intrinsèques de l'approche présentée ici. Le travail lexicographique ne peut reposer entièrement sur les corpus. Mais si les informations extraites de corpus doivent être contrôlées, corrigées, complétées, elles constituent néanmoins une vue d'ensemble sur l'emploi

[13] Selon G. Grefenstette, cet algorithme pourrait être modifié, éventuellement en exploitant une base de règles morphologiques de dérivations. La qualité des résultats devrait s'en trouver améliorée.

d'un mot et une source importante pour la rédaction d'entrées de dictionnaire. Pour exploiter ce type de données, le lexicographe devra acquérir l'expérience des outils permettant de les obtenir, afin de dépister les points faibles de telle entrée, identifier les associations douteuses, repérer les effets d'une analyse syntaxique inexacte ou ambiguë, et pour compléter les informations extraites par ses propres méthodes d'investigation.

3. FAIRE DES DISTINCTIONS DE SENS DE MOTS POUR LA RECHERCHE DOCUMENTAIRE

L'essor d'une société de la communication, avec notamment le développement d'un réseau donnant libre accès à de plus en plus de données textuelles, a profondément modifié les objectifs de la recherche documentaire. S'il s'agit toujours de sélectionner dans une base de documents un sous-ensemble de documents pertinents au regard des besoins d'un utilisateur, on a maintenant affaire à des bases approchant le milliard de mots (Evans et Zhai, 1996), où les textes de langue générale (ex. articles de presse) côtoient des textes de langue spécialisée relevant de domaines plus techniques.

3.1 Retrouver des textes dans une base documentaire

3.1.1 Principe général

Idéalement la requête de l'utilisateur spécifiant le type des documents recherchés devrait pouvoir être exprimée en langage naturel avec toute latitude dans le choix de la formulation ou, à la rigueur, dans un langage de requête, sous une forme explicite mais plus contrôlée. La formulation naturelle « *les textes décrivant les problèmes de circulation sur les grandes artères* » peut ainsi se traduire par une relation de localisation entre deux entités : LOCALISATION(problème de circulation, grandes artères). En pratique cependant, les systèmes commercialisés proposent généralement à l'utilisateur de formuler sa requête sous la forme d'une liste de mots clefs, éventuellement combinés par des opérateurs booléens (ex. circulation ET artères)[14].

Un système de recherche documentaire commence par indexer les documents de sa base, c'est-à-dire qu'il représente leur contenu sous la forme d'une liste de termes[15] représentatifs de ce contenu. Il extrait de la

[14] C'est ce type de requête qu'admet par exemple AltaVista, l'un des grands moteurs de recherche documentaire sur Internet. Il est accessible à l'adresse http://www.altavista.com.

[15] Dans le contexte de la recherche documentaire, le mot *terme* désigne une clé d'indexation.

même manière des termes de la requête de l'utilisateur. Puis, il cherche à apparier les termes de la requête avec ceux d'un document pour évaluer la pertinence de ce document au regard de cette requête. L'objectif est bien entendu de retrouver tous les documents pertinents de la base et ceux-là seulement. Dans la pratique, il faut trouver le meilleur compromis entre rappel et précision.

L'indexation est l'étape clef de ce processus de recherche documentaire. Comment représenter le contenu d'un document ? Les clefs d'indexation sont généralement des mots clefs : dans l'ensemble des mots d'un document, on sélectionne ceux que l'on suppose représenter le mieux le contenu du document, par exemple en éliminant les mots les plus fréquents et les moins fréquents supposés peu discriminants dans l'étape ultérieure de sélection des documents.

3.1.2 La question de la variation lexicale

Dans cette approche par mots clefs, qui est sans conteste robuste, se pose toutefois le problème de la variation lexicale. Considérons maintenant une deuxième requête, d'un étudiant en médecine : *« problème de circulation dans les artères ».* Un système fondé sur les mots clefs indexe cette requête comme celle de l'automobiliste mentionnée plus haut : (*circulation* ET *artère*). Il extrait donc le même ensemble de documents qui comporte aussi bien des textes sur la circulation sanguine que des textes sur la circulation automobile. En réponse à sa requête, l'automobiliste va donc trouver beaucoup de textes médicaux non pertinents pour lui (faible précision) tandis que des textes qui l'auraient intéressé ne sont pas sélectionnés parce qu'ils parlent de « *trafic* » et non de « *circulation »* (faible rappel). Prendre en compte les relations de synonymie (*trafic* / *circulation*) et de polysémie (*circulation sanguine* / *circulation automobile*) permettrait de gagner respectivement en rappel et en précision.

C'est généralement par une expansion de requête que l'on prend en compte les relations de synonymie autour des mots clefs de la requête. On enrichit la requête en indiquant quels synonymes peuvent être substitués aux mots clefs sans modifier le contenu de la requête : dans l'exemple ci-dessus on obtient ainsi la formule ((*circulation* OU *trafic*) ET (*artère* OU *axe*)). Cette expansion peut se faire soit automatiquement, soit sous le contrôle de l'utilisateur dans le cadre d'un système interactif qui l'aide à formuler sa requête en suggérant des synonymes.

Si la polysémie des mots de la requête peut également être traitée interactivement (le système peut de la même manière suggérer des distinctions de sens), pour réduire la polysémie dans les documents, il faut des méthodes de désambiguïsation automatique. Indexer un document non sur les mots clefs eux-mêmes (*circulation*) mais sur leur sens (*circulation* [*automobile*]) implique d'identifier le sens dans lequel le mot est employé dans un contexte donné.

Synonymie et polysémie sont en fait les deux faces du même problème : on voudrait fonder la recherche sur les sens de mots et non

sur les mots eux-mêmes. Dans le domaine très actif de la recherche documentaire, c'est l'un des axes qui est exploré. Sans développer les problèmes liés à l'expansion de requêtes (Voorhes, 1994), les paragraphes qui suivent mettent l'accent sur la désambiguïsation lexicale de gros volumes de textes tout-venant. S'il est trop tôt pour faire état d'expériences et de résultats sur des systèmes intégrant effectivement un traitement lexical, nous voudrions ici montrer l'une des pistes prometteuses, consistant à exploiter une base lexicale générale. Nous nous appuyons plus particulièrement sur le travail de M. Sussna (1993). Son impact sur un système de recherche d'information n'est pas réellement évalué mais il montre tout le parti qu'on peut tirer d'une base lexicale générale comme ***WordNet*** (voir chapitre III, *supra*).

3.2 Désambiguïser des corpus à l'aide de WordNet

M. Sussna (1993) défend l'idée qu'un système de recherche documentaire peut exploiter une source de connaissances comme ***WordNet*** pour désambiguïser des documents et les indexer sur les sens de mots plutôt que sur les mots. Son corpus d'expérimentation est un ensemble du *Time Magazine* comportant 425 articles de quelques centaines de mots en moyenne.

Les chapitres sur les corpus étiquetés et arborés ont montré les questions que soulève la désambiguïsation morpho-syntaxique ou syntaxique de corpus. Quelles informations morpho-syntaxiques ou quel niveau de structuration syntaxique faut-il représenter ? Comment assigner cette information aux différentes parties du corpus ? Ces questions se posent également pour la désambiguïsation lexicale. Quels sens de mots faut-il prendre en compte ? Comment identifier le sens d'un mot en contexte ?

Déterminer les sens à représenter pour un mot donné soulève en fait deux questions complémentaires. Celle de la granularité de la description : on peut retenir des distinctions de sens plus ou moins fines. Et celle des sources de connaissances : il s'agit de déterminer l'éventail des sens possibles pour un mot donné. M. Sussna (1993) propose d'exploiter les distinctions fines de sens telles que ***WordNet*** peut les représenter.

L'approche de M. Sussna est par ailleurs contextuelle. Comme beaucoup de travaux de désambiguïsation lexicale[16], elle repose sur l'idée que le contexte d'un mot permet d'identifier le sens dans lequel il est employé. Sous-jacente est l'intuition que l'on tend à sélectionner pour un mot le sens qui est lié au contexte. De fait, dans la plupart des cas, nous ne percevons pas d'ambiguïté car le contexte suffit à réduire l'espace des sens possibles. L'idée est de retenir pour un mot donné le sens qui se rapproche le plus de ceux de ses voisins, c'est-à-dire de mesurer la parenté ou la distance sémantique[17] entre les sens de différents mots qui

[16] Voir (Guthrie *et al*., 1994).
[17] Nous distinguons la notion de *parenté* sémantique de la mesure de *similarité*

se trouvent contigus dans le texte et de retenir la combinaison qui minimise la distance globale.

L'originalité de ce travail consiste à exploiter au maximum la structure de réseau de ***WordNet*** pour mesurer les distances entre les mots et à prendre en compte le problème de la co-détermination des sens dans une approche globale de la désambiguïsation. Nous développons ces deux aspects après avoir montré sur un exemple les résultats que M. Sussna cherche à obtenir.

3.2.1 Un article désambiguïsé

Sur un exemple d'article cité par M. Sussna (1993), nous montrons quel résultat peut être obtenu en exploitant les distinctions de sens de ***WordNet*** pour désambiguïser les sens de mots.

À partir de l'article original (point *a* ci-dessous), un premier traitement permet de sélectionner les « mots clefs » du document. Les noms étant traditionnellement supposés plus représentatifs du contenu d'un document que les autres catégories syntaxiques, M. Sussna ne conserve que les noms dans la représentation du document. Ceci suppose donc une étape de désambiguïsation morpho-syntaxique. On notera dans le résultat donné en *b* deux erreurs : *support* et *prime* ne sont pas employés comme noms dans l'article initial. En fait, M. Sussna ne retient que les noms présents dans ***WordNet***, ce qui élimine des noms propres (*Kennedy, MacMillan*) et des mots rares (*skybolt*) (point *c*). Il rejette de surcroît les mots réputés vides de sens et appartenant à un anti-dictionnaire (*stopword list*). Dans notre exemple, il s'agit de *december* mais surtout de noms propres très courants comme *U.S.*, *Europe*, *Europeans, Britain* (point *d*), à la différence de *France*. On obtient ainsi une liste de noms décrivant le contenu de l'article de départ (formule *e*).

C'est cette liste qu'il s'agit de désambiguïser en associant à chaque mot une étiquette spécifiant le sens dans lequel il est employé dans cet article. M. Sussna ne donne pas d'exemple de texte désambiguïsé mais nous proposons ci-dessous (point *f*) une version désambiguïsée de l'article *a*. Nous avons effectué cette désambiguïsation manuellement. Les étiquettes renvoient à des sens de ***WordNet*** (voir *supra* III.3). Le sens d'un mot est représenté par son numéro d'ordre dans la liste des sens possibles pour ce mot : c'est le 3^{e} des 6 sens de *strike* qui est employé ici. Ce sens est également décrit par le synset dans lequel il figure, *i.e*. l'ensemble de ses synonymes (entre accolades), ou à défaut, par la paraphrase (entre guillemets) donnée dans ***WordNet***[18].

***a*. texte original**

sémantique. La parenté, qui est généralement mesurée comme une distance entre les mots, peut recouvrir différents types de liens sémantiques : synonymie, antonymie, préférence sélectionnelle, y compris les relations de similarité qui mesurent plus spécifiquement un certain degré de substituabilité des mots en contexte (voir *supra* 3.2).

[18] Nous n'avons pas étiqueté ([sens = ?]) les mots qui ne sont pas employés comme noms et qui n'ont été conservés que du fait d'une erreur de catégorisation morpho-syntaxique. Nous ne donnons aucune description synonymique ou paraphrastique pour les noms qui n'admettent qu'un seul sens ([sens = 1/1]).

The allies after Nassau

In december 1960, the U.S. first proposed to help NATO develop its own nuclear strike force. But Europe made no attempt to devise a plan. Last week, as they studied the Nassau accord between President Kennedy and Prime Minister MacMillan, Europeans saw emerging the first outlines of the nuclear NATO that U.S. wants and will support. It all sprang from the anglo-U.S. crisis over cancellation of the bug-ridden skybolt missile, and the U.S. offer to supply Britain and France with the proved polaris (Time, dec. 28).

***b*. liste de noms**

allies Nassau december U.S. NATO strike force Europe attempt plan week Nassau accord President Kennedy Prime Minister MacMillan Europeans outlines NATO U.S. support crisis cancellation bug skybolt missile U.S. Britain France polaris

c*. liste de noms absents de *WordNet

Kennedy MacMillan skybolt

***d*. liste des noms figurant dans un anti-dictionnaire**

Nassau december U.S. NATO Europe Europeans Britain

***e*. liste de noms sélectionnés**

allies strike force attempt plan week accord president prime minister outlines support crisis cancellation bug missile france polaris time

***f*. liste de sens**

allies [sens = 1/3 : « an alliance of nations joining together to fight a common enemy »] strike [sens = 2/6 : « an attack that is intended to seize or inflict damage on or destroy an objective »] force [sens = 4/7 : {forcefulness, strength}] attempt [sens = 1/2 : {effort, endeavor, endeavour, try}] plan [sens = 1/3 : {program, programme}] week [sens = 3/3 : {calendar week}] accord [sens = 3/4 : {treaty, pact}] president [sens = 5/6 : {President of the United States, President, Chief Executive}] prime [sens = ?] minister [sens = 2/4 : {government minister}] outlines [3/3 : {schema}] support [sens = ?] crisis [sens = 2/2 : « a crucial stage or turning point in the course of something »] cancellation [sens = 1/2 : « the act of cancelling ; calling off some arrangement »] bug [2/5 : {glitch}] missile [sens = 1/2 : « a rocket-propelled weapon »] france [sens = 1/1] polaris [sens = 1/1] time [sens = 4/9 : « the continuum of experience in which events pass from the future through the present to the past »]

3.2.2 Mesurer la distance entre les nœuds de ***WordNet***

Pour M. Sussna, l'objectif est donc de mesurer par une distance entre les nœuds de ***WordNet*** la proximité des sens de différents mots dans un espace sémantique, c'est-à-dire leur parenté[19].

Traditionnellement, la distance de deux nœuds *a* et *b* dans un réseau est mesurée par la longueur du chemin le plus court entre *a* et *b*. Malheureusement, la taille de ***WordNet*** (cf. chapitre III, section 3.1.3) rend cette approche impraticable du fait du nombre de chemins à explorer pour calculer la distance entre deux nœuds.

Pour simplifier, on peut donc, comme le font E. Agirre et G. Rigau (1996) ou P. Resnik (1995b)[20], ne considérer que la partie hiérarchique de ***WordNet*** : « Soit *C* l'ensemble des concepts dans une taxonomie organisée autour de la relation EST-UNE-SORTE-DE (IS-A) telle qu'un nœud puisse hériter de plusieurs pères. Intuitivement, on peut considérer que deux concepts sont d'autant plus similaires qu'ils partagent plus d'information, cette information étant indiquée dans la taxonomie par le plus petit concept qui les domine tous les deux. La méthode reposant sur le décompte des arêtes mesure cela indirectement : si le chemin le plus court entre deux nœuds est tout de même long, cela signifie qu'il faut remonter haut dans la hiérarchie, jusqu'à des nœuds assez abstraits, pour trouver cet ancêtre commun. Par exemple, dans WordNet, NICKEL (*pièce de 10 cents en nickel*) et DIME (*pièce de 10 cents*) sont tous les deux dominés par COIN (*pièce*), alors que la classe la plus spécifique à laquelle appartiennent à la fois NICKEL et CREDIT-CARD (*carte de crédit*) est ASSET (*avoir*).»

Cette dernière méthode de calcul revient cependant à réduire ***WordNet*** à une hiérarchie de liens hyperonymiques et lui fait perdre une grande partie de sa richesse lexicale.

M. Sussna choisit de combiner ces deux approches du chemin le plus court et du chemin passant par le plus petit ancêtre commun. Il mesure la distance entre deux nœuds *a* et *b* par la longueur du chemin le plus court reliant *a* et *b* au sein de la sous-hiérarchie dominée par *p*, le plus petit ancêtre commun à *a* et *b* (figure 1, *infra*). Cette approximation paraît satisfaisante même si, parfois, on ne retrouve pas le chemin le plus court : dans le cas de la figure 1, le « raccourci » antonymique qui va de *a* à *b* en passant par *c*[21] est éliminé. Ce chemin peut être composé d'arêtes de différentes natures, liens hiérarchiques d'hyponymie, relations de méronymie, d'antonymie… Reprenons l'exemple de P. Resnik déjà cité au chapitre III (3.2.1). Le chemin *a* empruntant les liens hyponymiques de COIN à ASSET et de ASSET à CREDIT-CARD est de longueur 9, tandis que le

[19] Cette question du calcul de la distance sémantique se pose dans les mêmes termes, quelle que soit la source de connaissances exploitée. Plusieurs auteurs ont ainsi cherché à mesurer la parenté des sens de mots à partir de leur définition dans un dictionnaire et des mots qu'elles ont en commun. (Cowie *et al.*, 1992) et (Véronis et Ide, 1990), par exemple, exploitent respectivement le ***Longman Dictionary of Contemporary English*** et le ***Collins***.

[20] C'est nous qui donnons les équivalents français. Nous avons également remplacé MEDIUM-OF-EXCHANGE par ASSET pour rendre la citation cohérente avec la version 1.5 de ***WordNet*** et la figure ci-dessous qui s'en inspire.

[21] Les liens d'antonymie ne sont pas des liens hiérarchiques.

chemin *b* qui emprunte les liens hyponymiques de COIN à CURRENCY, le lien d'antonymie de CURRENCY à CREDIT et les liens hyponymiques de CREDIT à CREDIT-CARD est plus court (longueur 8). M. Sussna retient ce chemin qui est mixte mais plus court.

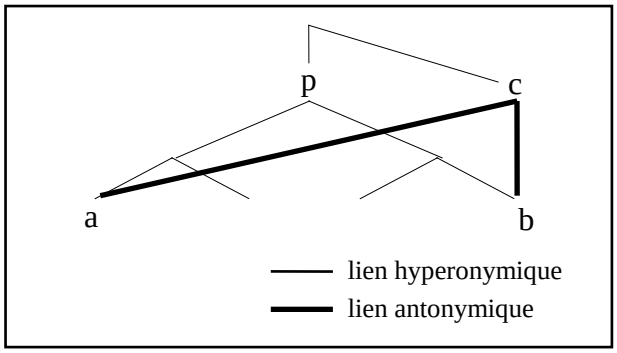

Figure 1.— Calcul du chemin le plus court au sein d'une sous-hiérarchie.

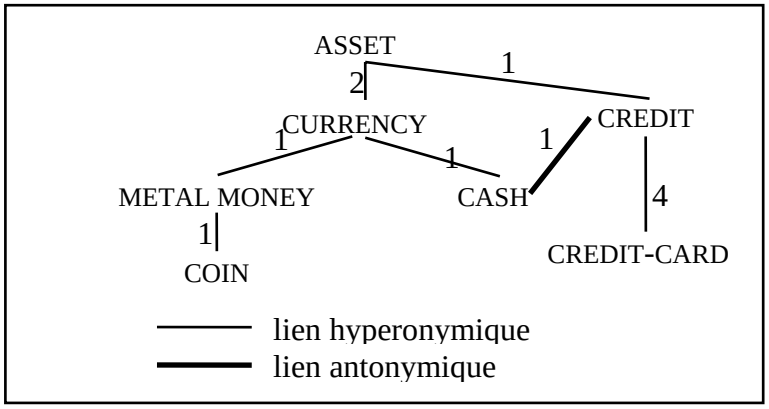

*Figure 2.— Calcul du chemin le plus court dans une sous-hiérarchie de **WordNet**. Pour aller de CREDIT-CARD à COIN, le chemin qui passe par le plus petit ancêtre commun (ASSET) est de longueur 9. Le chemin qui emprunte le lien antonymique entre CREDIT et CASH est plus court (longueur 8).*

Pour tenir compte de l'hétérogénéité des liens empruntés, M. Sussna pondère différemment chaque type de lien. Sans entrer dans le détail de ces poids qui sont déterminés expérimentalement, retenons les points suivants.

- Les liens de synonymie ont un poids nul et ne comptent pas dans les mesures de distance entre nœuds : les nœuds de ***WordNet*** étant des ensembles de synonymes (*synsets*), la synonymie est une relation interne aux nœuds.
- Les liens d'antonymie ont le poids le plus fort.
- Les poids des liens hyponymiques et méronymiques varient avec la « dilution » de la relation qui est mesurée en fonction du nombre de

liens de même type attachés aux nœuds concernés. Dans le cas, par exemple, de la relation A-POUR-PARTIE entre les nœuds VOITURE et PARE-BRISE, l'intuition est que cette relation reflète une parenté d'autant moins forte qu'une voiture comporte plus d'éléments (*i.e.* que plus de liens A-POUR-PARTIE partent du nœud VOITURE), mais d'autant plus forte, à l'inverse, que les pare-brises entrent dans la composition de moins d'objets (*i.e.* que moins de liens A-POUR-PARTIE arrivent au nœud PARE-BRISE). De fait le mot *pare-brise* évoque quasi automatiquement une voiture.

- Toutes les relations sont pondérées en fonction de leur profondeur dans la hiérarchie. Ce poids permet de tenir compte du fait que dans l'exemple de la figure 2 (*supra*), NICKEL et DIME sont plus proches que CREDIT et MEDIUM-OF-EXCHANGE, parce qu'ils sont situés plus bas dans la hiérarchie et reflètent donc des concepts plus spécifiques.

La longueur d'un chemin est donc calculée comme la somme des poids des différentes arêtes qui le composent et la distance entre deux nœuds est donnée par la longueur du chemin le plus court reliant ces deux nœuds au sein de la sous-hiérarchie dominée par le père commun.

C'est par l'expérimentation que M Sussna ajuste les différents paramètres de cette mesure. En ce qui concerne la diversité des liens à prendre en compte, M. Sussna montre, par exemple, en jouant sur les poids des différentes relations et en privilégiant les chemins hiérarchiques le long des liens hyponymiques, que l'on obtient de meilleurs résultats de désambiguïsation lorsqu'on exploite toute « la richesse des réseaux mixtes [comme WordNet], contenant à la fois des relations hiérarchiques et des relations non hiérarchiques » (Sussna, 1993). Les expériences menées par E. Agirre et G. Rigau (1996), qui donnent une « densité sémantique » dans ***WordNet*** comme mesure de la parenté entre les sens de mots, semblent montrer en revanche que les liens méronymiques apportent peu à la désambiguïsation[22]. Les conditions expérimentales et les mesures étant différentes, il est malheureusement difficile de comparer ces résultats.

Appréhender une parenté sémantique sous la forme d'une distance entre les sens de mots dans un réseau comme ***WordNet*** soulève ainsi de nombreuses questions. De multiples formules sont testées, mais il est encore beaucoup trop tôt pour tirer une conclusion définitive sur les paramètres à prendre en compte et pour se faire une véritable idée de leur impact sur les résultats de désambiguïsation. Seule l'expérience et le recul permettront de clarifier peu à peu cette question.

3.2.3 Désambiguïser un ensemble de mots

On peut donc désambiguïser un texte en retenant pour un mot donné le

[22] « *A priori*, une mesure de densité calculée à partir de relations plus nombreuses devrait d'autant mieux rendre compte de la notion de parenté sémantique et on pourrait s'attendre à de meilleurs résultats [de désambiguïsation] . Les expériences […] ont montré que la différence est négligeable ; ajouter l'information méronymique n'améliore pas la précision et n'augmente la couverture[22] que de 3% environ. » (*ibid*.) Ici, la couverture correspond à la proportion de noms effectivement désambiguïsés.

sens le plus proche des sens des mots voisins. M. Sussna propose une méthode de désambiguïsation globale qui respecte la co-détermination des sens. En effet, « si on ne calcule qu'un sens à la fois, comme le font la plupart des approches numériques de la désambiguïsation de mots, la question se pose de savoir s'il faut et comment on peut tenir compte du fait qu'un sens a été choisi pour un mot quand on cherche à désambiguïser le mot suivant ?» (Guthrie *et al*., 1994).

M. Sussna cherche à désambiguïser non pas un mot en fonction de son contexte mais un ensemble de mots conjointement en tenant compte de leur « contrainte mutuelle » (Sussna, 1993). Cela suppose de considérer toute la combinatoire des sens possibles, de calculer une distance binaire pour chaque couple de mot et de retenir la combinaison qui minimise la distance globale (l'énergie), somme des distances binaires. Le calcul de cette contrainte devient malheureusement vite prohibitif[23]. M. Sussna propose donc de désambiguïser conjointement les premiers mots d'un texte et de poursuivre « au fil du texte » en désambiguïsant chaque mot en fonction des sens retenus pour les mots qui le précèdent. Le contexte pris en compte dans le cas général est donc le seul contexte antérieur.

Pour déterminer la taille du contexte à considérer, M. Sussna procède, là encore, de manière expérimentale. En appliquant sa méthode à des fenêtres de tailles différentes et en comparant les résultats obtenus à une désambiguïsation aléatoire d'une part et à une désambiguïsation manuelle d'autre part, il constate que les résultats de la désambiguïsation s'améliorent quand on augmente la largeur de la fenêtre et se stabilisent pour une fenêtre de 41 mots. Sur ce point cependant, les expériences de (Agirre et Rigau, 1996) semblent montrer que la taille du contexte à prendre en compte dépend du type de corpus traité, les fenêtres réduites à 10 mots convenant pour le dialogue et les fenêtres plus larges donnant de meilleurs résultats pour les textes journalistiques.

3.3 De la désambiguïsation lexicale à la recherche documentaire

Si l'approche contextuelle de la désambiguïsation lexicale de corpus avait déjà été validée par différents travaux, le travail de M. Sussna montre le parti qu'on peut tirer d'un réseau comme ***WordNet***. La comparaison avec d'autres expériences montre cependant que le choix de la mesure de parenté sémantique, (le type des relations prises en compte, notamment) et le poids des conditions d'expérimentation (le type de corpus, par exemple) ont une grande influence sur les résultats. De la désambiguïsation lexicale à la recherche documentaire, un pas important reste à franchir.

[23] Pour une fenêtre de 10 mots et en ne retenant que 2 sens par mot, il faut déjà calculer 1 000 distances binaires, par exemple. Et si l'on considère la finesse des distinctions de sens faites dans ***WordNet*** et la sélection des noms retenus pour indexer un document, il faut compter avec beaucoup plus de sens par mot. A titre d'indication, dans la liste *f* donnée ci-dessus des noms décrivant le contenu d'un article de presse, les noms comportent en moyenne 3,7 sens.

Des questions plus fondamentales se posent par ailleurs. Elles concernent notamment la finesse des distinctions de sens à prendre en compte et la couverture des bases lexicales utilisées.

3.3.1 La granularité de la description lexicale

L'étiquetage décrit par M. Sussna est un étiquetage fin qui exploite les distinctions de sens de ***WordNet*** dans ce qu'elles ont de plus riche. Or on a vu au chapitre III que d'autres niveaux de distinctions de sens sont envisageables. Dans le cadre de la recherche documentaire, l'important est qu'il y ait correspondance entre la description de la requête et celle du document. Un compromis est à trouver entre la finesse de la description des sens et la capacité de l'utilisateur à préciser sa requête, à maîtriser ce niveau de description. On sait en effet que le commun des mortels ne maîtrise pas facilement toutes les distinctions de sens des lexicographes.

Si cette question de la granularité de la description n'est pas abordée par M. Sussna et il est encore difficile d'évaluer quel est le bon niveau de description pour la recherche documentaire.

3.3.2 La couverture des bases lexicales

L'exploitation de bases générales pour les tâches d'indexation pose un problème de la couverture. On a vu (chapitre III.1, *supra*) que les bases lexicales générales comme ***WordNet*** ne couvrent que partiellement les corpus spécialisés. Or les systèmes de recherche documentaire doivent indexer tout type de texte, des textes spécialisés comme des articles de presse. La question de la couverture est donc cruciale[24]. R. Krovetz (1991) indique que 50 à 60 % des mots susceptibles d'être retenus comme clefs d'indexation par un système de recherche documentaire sont absents du *Longman Dictionary of Contemporary English*. E. Agirre et G. Rigau (1996), qui travaillent sur un ensemble de textes diversifiés (différents types d'articles de presse, textes scientifiques et humoristiques), signalent que 11% des noms de leur corpus ne figurent pas dans ***WordNet***. C'est donc autant de mots qui ne peuvent pas être désambiguïsés.

Toute la question est donc de savoir quel intérêt peut avoir une désambiguïsation partielle pour un système de recherche documentaire.

Appréhender une parenté sémantique sous la forme d'une distance entre les sens de mots dans un réseau comme ***WordNet*** soulève ainsi plusieurs questions. De multiples formules de distance sont testées, mais il est prématuré de chercher à tirer une conclusion définitive sur les paramètres à prendre en compte et pour se faire une véritable idée de

[24] Si M. Sussna ne mentionne pas ce problème de couverture pour ***WordNet***, c'est probablement qu'il ne cherche à traiter que des articles de presse. En fait, c'est à dessein qu'il choisit ce corpus dans une base documentaire : « [n]ous travaillons à partir de la collection d'articles du Time Magazine qui est la moins spécialisée et la moins technique, parce que ***WordNet*** est un lexique de l'anglais général » (Sussna, 1993).

leur impact sur les résultats de désambiguïsation. Seule l'expérience et le recul permettront de clarifier peu à peu cette question.

4. UN MEME PARTI PRIS D'EMPIRISME

Ces travaux montrent que l'exploitation sémantique des corpus est largement empirique. Il s'agit toujours d'approcher le sens tel que le livre le corpus, en biaisant, à l'aide de techniques simples, souvent par une combinaison de techniques très spécifiques, chacune permettant de saisir un aspect particulier des phénomènes à décrire. Il en résulte une image imparfaite, souvent floue, mais qui néanmoins reflète le sens que l'on cherche à cerner. En retour, l'expérimentation devrait permettre de mieux comprendre les phénomènes observés.

4.1 Fonder une sémantique sur les corpus

Les expériences décrites ci-dessus témoignent d'un changement dans la vision même de ce qu'est la sémantique : on est passé d'une conception logique à une conception distributionnelle selon laquelle le sens d'un mot et plus largement d'une unité textuelle peut se décrire par les contextes dans lesquels il figure.

Au cours des années 1970 et 1980, c'est surtout l'Intelligence Artificielle qui s'intéresse à l'analyse sémantique de textes[25]. L'approche retenue est celle d'une compréhension en profondeur avec l'objectif de construire une représentation logico-sémantique, de la phrase, du paragraphe ou du texte. Il s'agit de modéliser les événements et situations dont parle le texte[26]. Mais, en dépit de leur intérêt théorique, la plupart de ces travaux n'ont pas été testés en vraie grandeur sur des textes réels, de plus d'une page, portant sur des domaines variés, comportant des mots inconnus et parfois mal rédigés, etc.

De même qu'en syntaxe, les techniques d'analyse robustes ont progressivement remplacé les techniques traditionnelles dans les systèmes destinés à traiter de gros volumes de textes tout-venant, de nouvelles approches sont aujourd'hui explorées pour l'analyse sémantique. Sous l'impulsion des besoins en matière de recherche d'information ou d'aide à la lexicographie spécialisée, l'objectif s'est déplacé. On ne cherche plus à comprendre tout le texte, à le représenter dans toute sa complexité, ses implicites et ses nuances de sens. Seule une partie du texte est pertinente, la représentation cible est généralement prédéfinie et on néglige les nuances de sens, les buts du locuteur, les

[25] Cf. (Herzog et Rollinger, 1991).
[26] Cela suppose tout à la fois de résoudre les anaphores, de repérer les variations de la prise en charge énonciative, de saisir la portée de telle négation ou de tel quantificateur, d'identifier les relations structurant l'ensemble du discours, etc.

présupposés et implicites, etc. L'accent porte désormais sur les problèmes de structuration lexicale avec notamment la désambiguïsation sémantique des mots, le calcul de contraintes de sélection, les phénomènes de synonymie, de parenté ou de classe sémantique et plus largement le repérage des relations lexicales.

Tous ces travaux reposent sur l'idée que le sens se construit en contexte mais aussi par le contexte. C'est donner un rôle central au corpus. On a souligné ce point dans le travail de G. Grefenstette. Celui de M. Sussna converge à et égard. Même lorsque des connaissances extérieures sont exploitées, elles n'ont pas le rôle que leur donnait l'Intelligence Articificielle. En introduisant des distinctions sémantiques supplémentaires, on peut caractériser plus précisément les contextes, mais c'est la confrontation des contextes entre eux qui fait émerger le sens. Les connaissances projetées sur le corpus ne servent alors que de révélateurs.

4.2 Exploiter des résultats approximatifs

Même si des perfectionnements sont envisageables, ces techniques sont approximatives. Les données ne sont jamais totalement fiables : la désambiguïsation des corpus reste imparfaite, un anti-dictionnaire n'est jamais ni complet ni totalement pertinent. Les opérations sont elles-mêmes approchées : l'extraction des fenêtres graphiques ne respecte pas totalement les frontières naturelles des zones textuelles (comme l'insertion d'un exemple ou d'une citation), le calcul des variantes morphologiques met l'accent sur le seul préfixe.

Le traitements effectués ne sont que partiellement maîtrisés. Par exemple, le volume des données à manipuler impose généralement de les « comprimer » : on élimine ainsi les mots outils, des mots trop rares, etc. Aucune de ces méthodes de compression de données n'est cependant neutre. Elles reviennent toujours à modifier la définition initiale du contexte et affectent les résultats. On a souvent souligné l'influence de la lemmatisation sur les performances de recherche documentaire (Church, 1995) et pour l'analyse de contenu Lebart et Salem, 1994) ou celle des mots fonctionnels (Riloff, 1995). Seule l'expérience pourra permettre de mesurer l'impact de ces traitements et d'ajuster les méthodes employées aux objectifs poursuivis.

Les résultats obtenus sont parcellaires. Souvent, seuls les noms sont pris en compte. Il y a plusieurs raisons à cela. La fiabilité des analyseurs ne permet pas toujours d'exploiter les contextes verbaux. La description lexicale des noms dans un réseau comme ***WordNet*** est plus riche et plus structurée — donc plus exploitable — que pour les autres catégories. Enfin, les techniques à mettre en œuvre ou les relations à exploiter diffèrent : on ne décrit pas un adjectif ou un verbe comme on décrit un nom. Pourtant, la description lexicale des adjectifs et des verbes est importante et des verbes peuvent être de bonnes clefs d'indexation (pour les corpus spécialisés notamment). Des méthodes ont été proposées pour

décrire les adjectifs ou les verbes[27], mais tout un travail d'expérimentation et de mise au point reste à faire pour construire automatiquement des ébauches d'entrées de dictionnaires effectivement exploitables. Quant à la question de la désambiguïsation des verbes, R. Basili et ses collègues (1997) soulignent qu'elle est peu explorée.

Les résultats obtenus sont néanmoins intéressants. Les entrées de dictionnaire construites automatiquement, même si elles demandent à être retravaillées par un lexicographe, donnent une vue globale du fonctionnement du mot dans un corpus technique. Elles aident à se repérer dans une langue spécialisée en s'affranchissant des préjugés induits par la langue générale. On peut supposer qu'une désambiguïsation lexicale même partielle augmente toujours la qualité de l'indexation d'un document et améliore la précision des systèmes de recherche documentaire.

4.3 Combiner des techniques simples

Les expériences rapportées ci-dessus reposent sur des techniques frustes au regard de l'ambition sémantique. Une fois données les ressources (corpus enrichi et/ou ressources lexicographiques générales), il s'agit d'extraire des contextes, de calculer des distances, d'éliminer les mots figurant dans un anti-dictionnaire, de comparer des préfixes de mots pour le calcul des variantes morphologiques, etc.. Aucune de ces opérations ne fait appel à un traitement sémantique, certaines ne nécessitent même aucune connaissance linguistique.

Dans la pratique, c'est souvent la combinaison de différentes techniques qui donne les meilleurs résultats. C'est patent dans (Grefenstette, 1993) qui fait appel à des techniques variées mais applique également une même technique, le calcul de similarités, sur des données de natures différentes. À chaque fois, une nouvelle facette du mot est mise en relief : les relations d'hyponymie dans lesquelles il entre, ses verbes opérateurs, les liens de parenté sémantique entre les mots. C'est en regroupant ces différentes informations qu'on peut construire des entrées de dictionnaires. Il faut également combiner différentes techniques pour la recherche de documents. Si l'on admet que l'indexation sur les sens plutôt que sur les mots améliore la précision de la recherche documentaire, il faut également cerner le rôle et la place de la désambiguïsation lexicale dans un système de recherche documentaire. Étant donnée la taille des bases documentaires à traiter, il est illusoire de chercher à désambiguïser et à indexer tous les documents au préalable. M. Sussna ne désambiguïse que des listes de mots présélectionnés. Il faut probablement aller plus loin et ne désambiguïser que certains textes ou certaines portions de textes qui auront été triés dans un premier temps par des techniques plus classiques de la recherche d'information (sur la

[27] Il s'agit de repérer le schéma de sous-catégorisation des verbes (Hindle, 1990 ; Resnik, 1993 ; Grishman et Sterling 1994) ou les liens d'antonymie et les relations scalaires entre les adjectifs (Justeson et Katz, 1996 ; Hatzivassiloglou et MacKeown, 1993).

base de mots clefs statistiquement significatifs, par exemple).

Plus généralement, il s'agit de trouver le bon dosage des méthodes linguistiques et statistiques. (Sussna, 1993) semble postuler que la description la plus riche est nécessairement la plus appropriée. Cela ne va pas de soi. Nous avons vu que des distinctions fines de sens peuvent n'être pas pertinentes pour la sélection de documents (voir *supra*, 3.3.1). De la même manière, il n'est pas certain que la lemmatisation systématique (Church, 1995) ou la morphologie dérivationnelle, avec notamment le regroupement des mots appartenant à la même famille dérivationnelle (*stemming*) (Gaussier et coll., 1997), améliore les performances de la recherche documentaire. Par ailleurs, le travail de G. Grefenstette (1993) le montre, les traitements linguistiques sont lourds et peuvent souvent être convenablement approchés — parfois supplantés — par des techniques frustes.

4.4 Modéliser par ajustements successifs

C'est toujours de manière empirique qu'on cherche à rendre compte du sens que véhicule le texte. On tente de construire un modèle qui décrive au mieux les effets de sens observés ou perçus. Ce modèle n'est pas construit *a priori,* il est progressivement mis au point au vu des résultats obtenus. Ce travail d'ajustement permet en retour de mieux comprendre la nature des phénomènes décrits.

Le volume des textes à traiter impose de s'affranchir du détail de tel effet de sens et de la diversité des phénomènes de surface pour donner une description synthétique du corpus. Dans les exemples présentés ici, comme souvent, cette modélisation repose sur des mesures quantitatives et statistiques. La mesure, en effet, même si elle a peu de signification en tant que telle, permet de résumer un ensemble d'observations, de comparer et d'ordonner les phénomènes observés.

La démarche consiste généralement à emprunter un modèle connu dont les propriétés ont le mérite d'être bien décrites puis à en ajuster expérimentalement les paramètres pour affiner la description et mieux rendre compte des phénomènes perçus. On cherche ainsi à approcher la notion de parenté sémantique par des mesures de distance vectorielle ou de distance dans un graphe. Diverses expériences ont été menées pour modéliser l'opération de désambiguïsation sémantique à l'aide d'un réseau de neurones (Véronis et Ide, 1990) ou par la méthode du recuit simulé empruntée à l'algorithmique combinatoire (Cowie *et al.*, 1992). Il reste ensuite à ajuster le modèle en modifiant le nombre ou la nature des paramètres pris en compte et en jouant sur leurs poids respectifs. C'est par une série d'expériences que M. Sussna détermine la taille des contextes et le poids de chaque type de relation dans le calcul de la distance sémantique des nœuds de ***WordNet***. Après avoir « testé une grande variété de mesures de similarités » entre les mots, G. Grefenstette retient celle « qui semble produire les meilleurs résultats » (1994a, p. 47).

Il n'est donc pas de « bon » modèle dans l'absolu. Il n'existe que des

modèles opératoires qui sont utiles à l'utilisateur final dans le cadre d'une application donnée. Seul le lexicographe peut dire si les ébauches d'entrées de dictionnaires construites automatiquement lui fournissent effectivement un bon point de départ. C'est dans la mesure où la désambiguïsation lexicale telle que l'envisage M. Sussna permet d'améliorer significativement la précision de la recherche de documents qu'elle présente un intérêt, par exemple. Le verdict d'utilité est la seule véritable évaluation possible. La maturité du domaine ne permet malheureusement pas toujours de mener cette évaluation globale à bien, mais l'exemple des entrées de dictionnaire construites par SEXTANT montre néanmoins la fécondité de cette démarche empirique.

En ce qui concerne l'étiquetage morpho-syntaxique et syntaxique, il existe des corpus étiquetés qui font l'objet d'un consensus suffisant pour servir de référence et on peut comparer entre eux les résultats obtenus par des méthodes différentes. En matière sémantique, en revanche, la subjectivité des phénomènes et la diversité des objectifs se traduisent par une grande hétérogénéité des étiquetages et interdisent toute évaluation intermédiaire.

4.5 Expérimenter pour mieux expliquer

Toute la difficulté vient de qu'en modélisant, on cherche à rendre compte de notions qui sont essentiellement intuitives et largement subjectives. Pour un locuteur donné, la notion de parenté sémantique repose sur des associations d'idées toutes personnelles et on sait que la définition d'un mot varie d'un dictionnaire à l'autre, y compris pour ce qui est de la distinction de ses différents sens.

On arrive ainsi à un paradoxe. On observe l'extrême sensibilité des résultats au mode de calcul utilisé, aux paramètres pris en compte et à leurs poids respectifs. Par des réglages expérimentaux, on sait construire des modèles opératoires qui décrivent effectivement les effets de sens dans un corpus donné. Pour autant, on ne sait pas toujours expliquer pourquoi tel modèle est meilleur que tel autre.

Pourtant, ces expériences devraient progressivement permettre de mieux comprendre en retour les phénomènes que l'on cherche à modéliser. La diversité des conditions expérimentales fait qu'il est souvent difficile de tirer des conclusions générales sur les propriétés de telle mesure, l'importance de tel paramètre ou l'adéquation de tel modèle et nos connaissances en la matière sont encore parcellaires et fragiles. Pourtant, l'expérimentation systématique consistant à tester un à un différents paramètres comme le font M. Sussna (1993) ou G. Grefenstette (1994a), la confrontation de différentes mesures sur les mêmes données expérimentales, comme le fait (Daille, 1994) par exemple, commencent à porter leurs fruits. La convergence des résultats de différents auteurs (Sussna, 1993 ; Agirre et Rigau, 1996 ; Resnik, 1995b) montre que la parenté sémantique d'un ensemble de mots est

perçue comme d'autant plus grande que leurs sens sont plus précis[28].

Le cas du score d'association est exemplaire de cette démarche empirique. K. Church et P. Hanks ont proposé (1990) de mesurer la force de cooccurrence de deux mots par une mesure fondée sur la notion d'information mutuelle et empruntée à la théorie de l'information. Ils ont montré l'intérêt et la diversité des résultats qu'elles permettait d'obtenir. À leur suite, de nombreux auteurs ont eu recours à cette mesure (Hindle, 1990 ; Resnik, 1995b). Pourtant le choix de cette mesure n'est jamais réellement justifié : on en explicite les propriétés formelles, mais sans expliquer pourquoi cette mesure est pertinente pour mesurer des contraintes de sélection. La convergence de différentes expériences montre cependant qu'en donnant un poids important aux événements rares et en soulignant les emplois « spécialisés »[29], le score de cooccurrence fait ressortir les expressions figées, ce qui est précieux dans une perspective lexicographique : l'association de *œil* et de *boeuf*, dans *oeil de bœuf*, est intéressante pour la description du mot *boeuf*. Mais ceci explique à l'inverse que cette mesure soit mal adaptée à la modélisation conceptuelle d'un domaine, ce que (Habert *et al.*, 1996) met en évidence. Pour décrire le concept auquel renvoie un mot, ses propriétés et les relations dans lequel il entre, il faut au contraire éliminer les attirances proprement lexicales et s'appuyer davantage sur les associations banales comme *manger/élever du bœuf*, *viande de bœuf*, *bœuf cuit*, etc. L'information mutuelle est donc un bon indice lexicographique mais un mauvais outil de modélisation conceptuelle. Par ailleurs, cette mesure qui « met l'accent sur les phénomènes rares » (Basili *et al*., 1993b, p. 179) est peu adaptée aux contextes syntaxiques : elle serait utile « si on pouvait se fier entièrement aux analyses » (*ibid*.), mais elle donne en fait trop d'importance à des relations « dues à des ambiguïtés syntaxiques ou des erreurs d'analyse » (*ibid*.). C'est la multiplication et la confrontation des expériences utilisant la mesure de l'information mutuelle et la comparaison avec des mesures différentes qui permet de tirer des conclusions de portée un peu générale, de progressivement mieux comprendre ses propriétés comme mesure de distance entre les mots et de cerner les conditions de son utilisation.

[28] Pour un sens donné, on peut mesurer ce degré de spécificité ou « *contenu informationnel* » (Resnik, 1995b) par la hauteur du nœud qui le représente dans une hiérarchie comme ***WordNet*** ou par le nombre de nœud que ce nœud domine.
[29] Le fait pour un mot de figurer toujours ou très souvent dans le(s) même(s) contexte(s).