Classer les mots : sémantique à gros grain et méthodologie harrissienne[*]

Benoît Habert[*] — Pierre Zweigenbaum[**]

[*] LIMSI – Groupe LIR (Langues, Information et Représentations)
BP 133
F–91403 Orsay Cédex
habert@limsi.fr

[**] Mission de Recherche en Sciences et Technologies de l'Information Médicale,
DSI, Assistance Publique - Hôpitaux de Paris

STIM – CHU Pitié-Salpêtrière
91, boulevard de l'Hôpital
F–75634 Paris Cedex 13
pz@biomath.jussieu.fr

Résumé

La mise en regard de la démarche de découverte de classes d'opérateurs et d'opérandes dans les derniers travaux de Harris, à la fin des années quatre-vingts et des méthodes d'acquisition sémantique automatique, en plein essor dans la dernière décennie en traitement du langage naturel et en recherche d'information, met en évidence les angles morts de chaque approche et leur complémentarité possible.

Abstract

The current semantic category acquisition methods developed in Natural Language Processing and Information Retrieval and the discovery procedures for semantic sublanguage grammars that Harris devised in the late eigthies are being compared. This comparison yields an appraisal of each framework. It leads to cooperation prospects.

1 Une sémantique à gros grain face aux « déferlantes » (hyper)textuelles

Le nombre de pages Web accessibles via la Toile dépasse les trois milliards. S'ajoutent les nombreuses et vastes bases textuelles et documentaires d'accès plus restreint (intranets, par exemple). Le contraste est violent entre cette masse (hyper)textuelle et la « rusticité » des moyens dont tout un chacun dispose pour y accéder. Les moteurs de recherche écrèment les documents qu'ils indexent comme les requêtes qui leur sont soumises. Les mots « vides » (les mots outils) sont enlevés. Les mots restants sont *racinisés*, sans lemmatisation morphologique, mais par troncation de terminaisons, pour faciliter les rapprochements (*sémantique* pourra être ramené à *sémant*, ce qui permettra de rapprocher les documents comprenant ce mot de ceux contenant *sémantisme*, *sémanticien*, *sémantème*...). Une importance plus grande est accordée aux mots qui permettent de discriminer entre les documents, précisément parce qu'ils ne sont pas présents dans tous les documents. Les documents rapportés par un moteur de recherche pour une requête donnée sont alors ceux dont la représentation simplifiée et « pondérée » est la plus proche de la requête, elle-même simplifiée et pondérée. Ils sont alors généralement classés par proximité décroissante avec la requête. Un examen exhaustif des résultats d'une requête permettrait d'évaluer la *précision* du moteur utilisé, c'est-à-dire la proportion de documents pertinents dans l'ensemble des documents rapportés (le complémentaire de la précision est le *bruit*). L'homonymie et la polysémie

[*] Nous reprenons, dans un ordre différent et avec un angle de vue changé, l'essentiel de la matière de (Habert & Zweigenbaum, 2002).

contribuent au bruit : c'est ainsi qu'une requête sur *Charles Fillmore* permet certes d'accéder aux documents concernant le linguiste, mais l'essentiel des pages fournies concerne un prêcheur télévisuel à succès du même nom. Par contre, puisqu'il est matériellement impossible d'examiner l'ensemble des documents indexés par le moteur de recherche, on n'est pas en mesure d'évaluer le *rappel*, c'est-à-dire la proportion de documents pertinents effectivement ramenés par rapport à l'ensemble des documents pertinents (le complémentaire du rappel est le *silence*). La non-prise en compte des synonymes engendre du silence : les documents qui contiennent uniquement des équivalents des mots de la requête restent dans l'ombre.

Bruit et silence brouillent notre accès aux données (hyper)textuelles. C'est pour cette raison que le consortium qui gère la Toile (http://www.w3.org) a lancé le projet d'un Web sémantique : l'accès aux textes par le sens. Ce projet fédère au niveau mondial de très nombreuses initiatives. Il correspond à plusieurs directions de travail. La désambiguïsation sémantique (*WSD – Word Sense Disambiguation*) tout d'abord : étant donné un inventaire de sens donné pour un mot, il s'agit d'être capable de trouver l'acception effective en contexte. Ide & Véronis (1998) fournissent un état de l'art du domaine. L'acquisition sémantique automatique ensuite : la constitution ou la complétion, automatique ou assistée, de dictionnaires sémantiques, soit généraux, soit spécialisés. C'est sur cette tâche de classement sémantique de mots que nous centrons ces pages.

Précisons d'emblée les traits principaux de cette sémantique des textes électroniques. Elle s'attache presque exclusivement au lexique et en particulier aux relations lexicales (synonymie, homonymie, polysémie, mais aussi hyperonymie et méronymie). Elle vise la robustesse : la possibilité de traiter du texte « tout venant » , « révisé » ou non, en quantité quelconque. C'est une sémantique sinon mesurée, du moins « mesurante » : la quantification y est centrale et la logique absente. La tradition de sémantique proprement linguistique y pèse moins que l'ancrage dans la tradition philosophique des ontologies (Sowa, 2000) : sont visés moins les sens que les « concepts » dénotés par les mots, ce qui ne va pas sans naïvetés (l'illusion de tenir les choses par les mots).

Nous présentons d'abord les méthodes de l'acquisition automatique (section 2). Nous revenons ensuite sur les propositions, proches, faites par Harris dans ses derniers travaux (section 3). La mise en regard des deux familles de travaux permet enfin un début d'évaluation de cette sémantique à gros grain (section 4) et ouvre des perspectives de coopération (section 5).

2 Acquisition sémantique automatique : un survol

Classer des mots est une dénomination ambiguë, qui recouvre en fait deux activités distinctes, mais souvent complémentaires, partitionner et répartir. Dans le premier cas, on veut regrouper les mots « qui se ressemblent » entre eux (par ensemble de synonymes, par exemple). Le résultat est une partition des mots. On parle alors de *classification* (*clustering*). Dans le second cas, on dispose de catégories pré-déterminées, d'un *répartitoire*, pour reprendre la terminologie de Damourette et Pichon : il s'agit de placer chaque mot dans la catégorie qui lui convient. On parle alors de *catégorisation* (*categorization*). L'opposition et l'interaction entre les deux activités est analogue à celles de la botanique et de la zoologie entre *mise au point de taxonomies* (organiser les êtres vivants en espèces, genres...) et *détermination* ou *identification* (trouver l'espèce, le genre, etc., dont relève une plante ou un animal). En apprentissage automatique (Mitchell, 1997), c'est l'opposition entre un *apprentissage non supervisé*, qui ne connaît pas *a priori* les classes qui peuvent être trouvées et un *apprentissage supervisé*, où les catégories pré-existent. On notera incidemment que l'activité d'un moteur de recherche est un cas particulier de répartition, de catégorisation : il s'agit en effet de positionner un document plus ou moins près d'une des deux catégories pré-existantes, pertinent et non pertinent.

2.1 Partitionner des mots en « classes »

L'objectif global est de trouver des « airs de famille » entre les mots. A partir du moment où l'on ne saurait prendre en compte toutes les caractéristiques de ces mots, a fortiori puisqu'il importe d'obtenir une méthode automatisable, la première étape est celle de la réduction, de la simplification des traits associés

aux mots. C'est une optique distributionnelle qui prévaut : un mot est caractérisé par les contextes dans lesquels il figure. La deuxième étape est celle de l'obtention d'un indice synthétique de la proximité relative entre deux mots. Elle se concrétise par le calcul d'une *distance* entre les mots deux à deux en fonction des contextes qu'ils partagent, de ceux qui sont propres au premier, de ceux qui sont propres au second et de ceux qui ne sont employés ni par l'un ni par l'autre. La distance de Jaccard se calcule par exemple ainsi :

$$\frac{|\,contextes\ partagés\,|}{|\,contextes\ partagés\,|+|\,contextes\ propres\ à\ mot_1\,|+|\,contextes\ propres\ à\ mot_2\,|}$$

Rappelons que les barres verticales notent la cardinalité d'un ensemble. Lorsque deux mots partagent tous leurs contextes, la proximité est de 1, lorsqu'ils n'en partagent aucun, elle est de 0. La troisième étape consiste à regrouper les mots en sous-ensembles en fonction des distances obtenues à l'étape précédente.

Chacune de ces trois étapes peut donner lieu à des choix différents (on reviendra en section 4 sur les conséquences de ces choix). Tout d'abord, la définition du contexte peut être syntaxique. On peut par exemple associer à un nom les adjectifs qui le modifient, les verbes dont il est le sujet, ceux dont il est l'objet. Le contexte d'un mot peut également être défini comme une « fenêtre » dans laquelle figure ce mot-pivot. La taille d'une fenêtre peut dans ce cas varier, soit en prenant des limites structurelles (la phrase, le paragraphe...) soit en prenant un nombre déterminé de mots avant et après le mot-pivot. Au sein de cette fenêtre, on peut éventuellement éliminer les mots-outils. On peut également faire appel à un programme d'étiquetage et de lemmatisation. En second lieu, de nombreuses mesures de distance ont été mises au point (Losee, 1998, p. 43–62). Enfin, s'offrent différentes techniques de classification (Lebart *et al.*, 1997, p. 145–185), certaines opérant de manière descendante, par séparations successives, d'autres travaillant de façon ascendante, par agglomérations progressives.

Le tableau 2.1 montre quelques contextes possibles pour un mot-pivot dans une phrase extraite d'un corpus de notations d'infirmières sur des très grands prématurés. La figure 1 fournit les regroupements des 50 adjectifs les plus fréquents de ce corpus. La distance retenue est celle de Jaccard. Le contexte est limité aux lemmes des mots-pleins dans une fenêtre de 5 mots avant et après chaque adjectif. La classification employée est ascendante : vers le bas de l'arbre résultant, les mots ont des contextes très proches (par exemple *endormi* et *fatigué*. Au fur et à mesure qu'on se rapproche de la racine de l'arbre (en haut du graphique), les proximités s'amuissent. Une telle classification fournit des (propositions de) « classes » emboîtées les unes dans les autres. Par exemple, le rapprochement *endormi-fatigué* et la proximité *hypotonique-douloureux* s'agglomèrent pour former un nouveau regroupement. On peut alors « couper » l'arbre à un niveau donné pour obtenir des classes du « grain souhaité.

Tableau 1 : *Variation de contextes autour d'un 'pivot'*

Phrase	
ce bébé **agrippe** *très bien le doigt quand on le lui met dans la paume de la main*	
Contexte syntaxique	
Phrase entière, tête des constituants	{bébé-Sujet}{doigt-Objet}
Fenêtre graphique	
4 mots autour du pivot	{ce, bébé, très, bien, le, doigt}
+ étiquetage	{ce-Dét, bébé-N}{très-Adv, bien-Adv, le-Dét, doigt-N}
+ mots pleins seulement	{bébé}{doigt}

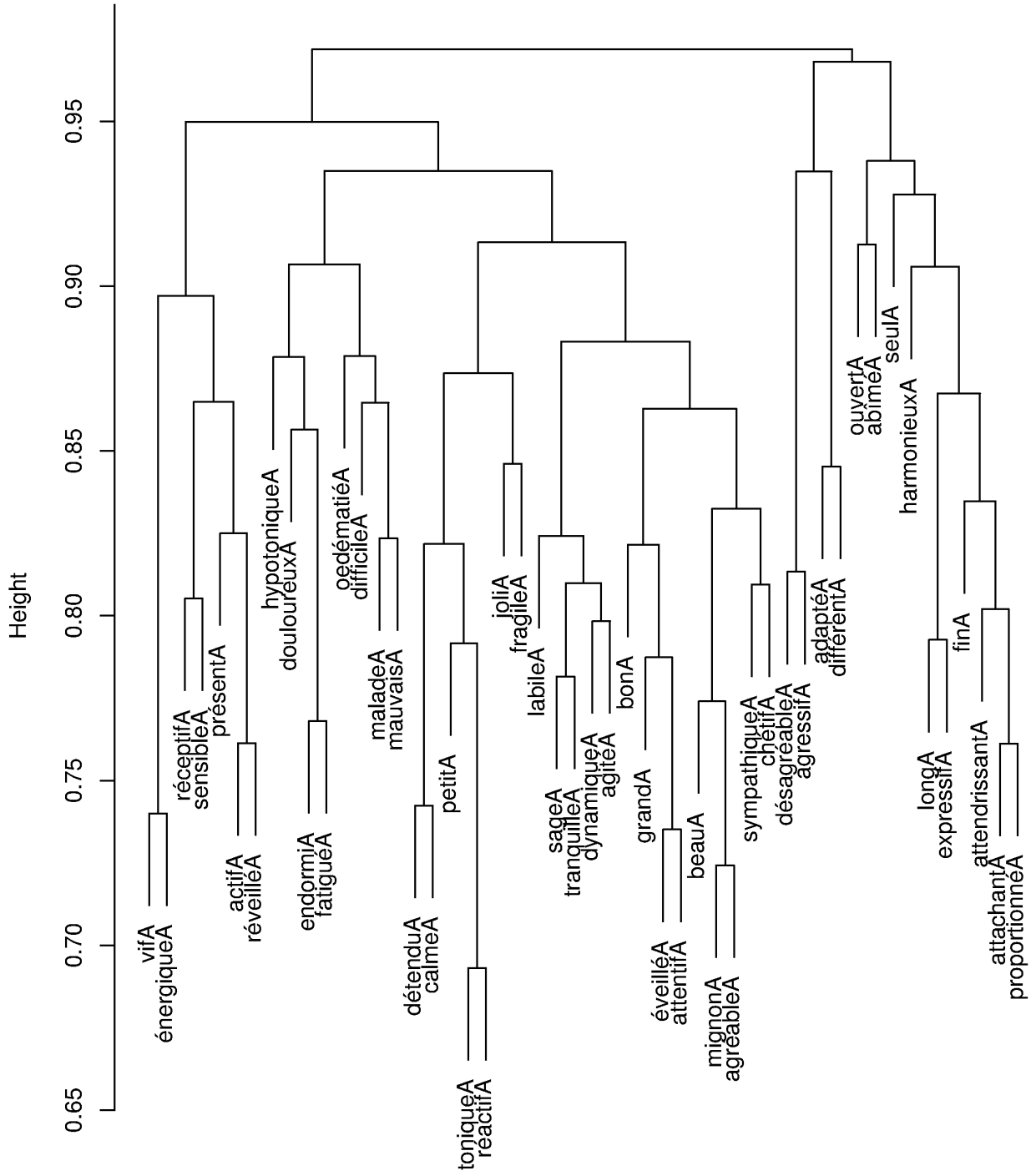

Figure 1 : Regroupement des 50 adjectifs les plus fréquents. Corpus Prématurés 00 – f : 5 lemmes dans phrase, Jaccard.

2.2 Répartir des mots dans des classes

Les similarités de contextes entre mots, base des partitionnements de mots en classes, peuvent également servir à répartir des mots dans des classes pré-existantes. Etant donné un ensemble de mots assortis d'une étiquette sémantique, on peut chercher quels sont les *k* mots étiquetés les plus proches d'un

mot non étiquété donné. Si ces *k* « voisins » portent tous la même étiquette sémantique, on peut donner cette étiquette au mot examiné. Si ces *k* voisins se partagent entre plusieurs étiquettes, on peut attribuer au mot non étiqueté la catégorie qui a la majorité. C'est l'approche retenue par Nazarenko *et al.* (2001).

Une autre approche s'appuie sur des « patrons » lexico-syntaxiques caractéristiques d'une relation sémantique donnée, comme ceux indices de l'hyponymie en français mis en évidence par Borillo (1996). Un patron du type *SN tel que SN+*, c'est-à-dire un syntagme nominal suivi de *tel* (ou *telle*, *tels*, *telles*), de *que* et d'un ou de plusieurs groupes nominaux correspond souvent à une relation d'hyperonymie entre le premier SN et ceux qui suivent *tel que*, comme dans la phrase *Des cations tels que le sodium, le potassium, le calcium et le magnésium peuvent être dosés par une méthode de routine*. Si l'on connaît la classe sémantique de *cations*, on peut l'attribuer à *sodium*, *potassium*, *calcium* et *magnésium*. Cette approche, présentée par Hearst (1992), a été largement appliquée et améliorée. Le repérage « manuel » de tels patrons fait désormais place à des techniques d'apprentissage automatique de règles, qui tablent sur les distributions observées des mots d'une même classe sémantique (Morin, 1999).

Deux paramètres interviennent dans cette utilisation d'un répartitoire prédéfini. Le premier est le nombre de mots déjà étiquetés. Il peut s'agir de lexiques sémantiques d'une taille proche de celle de dictionnaires de langue (une centaine de milliers de vedettes). Mais il peut s'agir aussi d'une poignée de mots pour chacune des classes du répartitoire. On parle dans ce cas-là de mots-amorces (*seed words*). Le deuxième paramètre est le caractère général ou spécialisé du répartitoire : les classes peuvent être générales ou au contraire correspondre à un domaine d'application étroit.

3 Le dernier Harris : une méthodologie pour découvrir des grammaires sémantiques

L'utilisation depuis une dizaine d'années de méthodes globalement distributionnelles en acquisition sémantique automatique (Grefenstette, 1994a) s'est accompagnée d'un « retour à Harris » , par exemple dans (Church *et al.*, 1991 ; Brill & Marcus, 1992). Ce retour à Harris privilégie paradoxalement le distributionnalisme des années 50–60. Pourtant, les derniers travaux de Harris, à la fin des années quatre-vingts – cf. (Daladier, 1990 ; Ryckman, 1990 ; Dachelet, 1994) pour une introduction en français – proposent une méthodologie syntaxique précise pour découvrir des classes sémantiques de mots et les patrons syntactico-sémantiques dans lesquelles elles figurent. Harris table sur une distinction entre langue générale et *sous-langages*. Il est instructif de mettre en rapport cette distinction et de cette méthodologie d'un côté et l'acquisition sémantique automatique actuelle.

Nous présentons dans la section 3.1 les hypothèses fondatrices : les relations de sélection, conduisant à des distinctions de sens, peuvent être mises au jour de manière objective ; leurs frontières sont clairement déterminées pour les sous-langages mais floues pour la langue en général. L'analyse syntaxique d'un corpus pertinent et la normalisation des relations syntaxiques sous-jacentes permettent de dégager des catégories et des patrons sémantiques (section 3.2). Ces catégories et ces patrons reflètent le monde perçu, son évolution, et les relations qui l'organisent (section 3.3). Nous laissons à dessein les citations de Harris en américain, en particulier pour rendre plus lisible la distinction entre *meaning* et *information* (Nevin, 1993 ; Leeman, 1996).

3.1 Hypothèses fondamentales

Le sens : un résultat, pas un point de départ

Pour Harris (Harris, 1988, p. 60), *... there is no usable classification and structure of meanings* per se, *such that we could assign the words of a given language to an* a priori *organization of meaning*. Un exemple frappant le souligne (*ibid*., p. 62) : *The operator* divide *has virtually the same meaning as the operator* multiply *when its argument is a cell name* [le nom d'une cellule du corps]*: for a cell, to divide is to multiply*. Il n'est dans ces conditions pas possible de se fonder sur un sens *a priori* des mots. Pour Harris, il n'existe pas, pour analyser le langage, de métalangue externe à la langue. Si bien que *[p]our spécifier comment la langue "véhicule" l'information, la grammaire n'a pas la ressource de réduire la*

langue à un précédent "message" ou "langage interne" ou à quelque chose de non linguistique tels que objets, événements ou propriétés du "monde réel" (Ryckman, 1990, p. 25).

Cette position ne conduit pas Harris à abandonner pour autant toute investigation de type sémantique. Il soutient au contraire que les relations de dépendance entre un mot et les opérandes dont il dépend ou les opérateurs qui dépendent de lui sont *objectively investigable and explicitly statable and subdividable* (Harris, 1991, p. 332) et conduisent à des distinctions sémantiques : *Characterizing words by their selection allows for considering the kind and degree of overlap, inclusion, and difference between words in respect to their selection sets — something which might lead to syntax-based semantic graphs (e.g. in kinship terms), and even to possibilities of decomposition (factoring) for particular sets of words. Such structurings in the set of words are possible because in most cases the selection of a word includes one or more coherent ranges of selection ... The effect of the coherent ranges is that there is a clustering of particular operators around clusterings of particular arguments, somewhat as in the sociometric clusterings of acquaintance and status (e.g. in charting who visits who within a community)* (*ibid*., p. 329–330). L'information découle alors de ces relations de dépendance : *[Harris] a développé l'argument que les hiérarchies de sélection sur les combinaisons d'éléments linguistiques comprennent ce qu'on peut considérer comme la structure informative d'une phrase ou d'un texte* (Ryckman, 1990, p. 25).

On pourrait vouloir rapprocher cette position de la citation de Firth sur les collocations – *you shall know a word by the company it keeps* (Firth, 1957), rendue en particulier fameuse par l'utilisation de l'*information mutuelle* pour mesurer l'attirance entre mots (Church *et al.*, 1991) ; (Manning & Schütze, 1999, ch. 5) et pour proposer des regroupements et éventuellement des classes sémantiques « grossières » sur la base des cooccurrents isolés par l'information mutuelle et partagés (Church & Hanks, 1990). En fait, et c'est un point de divergence important avec une bonne partie des techniques de partitionnement automatique des mots présentées supra, Harris met l'accent sur des dépendances (récursives) et non sur des cooccurrences (fréquentes) : *... the structural property is not merely co-occurrence, or even frequent co-occurrence, but rather dependence of a word on a set: an operator does not appear in a sentence unless a word — one or another — of its argument set is there (or has been zeroed there). When that relation is satisfied in a word-sequence, the words constitute a sentence* (*ibid*. p. 332). C'est ce rapport de dépendance qui rapproche les mots : *The word classes are ... defined by their dependence on word classes which are in turn defined by the same dependence relation* (*ibid*., p. 17).

Vraisemblance

« Chaque mot a une probabilité spécifique et relativement stable d'apparaître comme argument, ou opérateur, avec un autre mot, même si l'on rencontre de nombreux cas d'incertitude, de désaccord entre les locuteurs et de changement au cours du temps » : c'est la contrainte de *vraisemblance* [*likelihood*] de Harris (1988), rappelée dans (Pereira, 2000, pp. 1241–1242). Cette vraisemblance peut ainsi être vue comme une probabilisation des événements linguistiques, ici, la dépendance entre un opérateur et un opérande particulier, c'est-à-dire des restrictions de sélection. La forte probabilité d'un élément est aussi un point clé de la contrainte de *réduction* (*ibid.*) : « Elle consiste, pour chaque langue, en quelques types spécifiables de réductions... ce qui est réduit... est le matériau hautement probable...un exemple est l'effacement [*zeroing*] des mots correspondants répétés sous *et*. » Pour résumer, *... it is an essential property of language that the combinations of words in utterances are not equiprobable, and in point of fact that many combinations do not appear at all* (Harris, 1991).

Caractérisation probabiliste vs. *booléenne des sélections : de la langue aux sous-langages*

Les régularités sélectionnelles diffèrent cependant pour les sous-langages par rapport à la langue. Naomi Sager, qui a travaillé avec Harris, fournit la définition suivante : *Informally, we can define a sublanguage as the language used by a particular community of speakers, say, those concerned with a particular subject matter or those engaged in a specialized occupation* (Sager, 1986, p. 2). Pour elle (*ibid.*, p. 3), comme pour Harris, la sélection des opérandes dépendant d'un mot ou des opérateurs gouvernant un mot est de type probabiliste pour la langue (*langage as a whole*, écrit Harris) et de type booléen pour les

sous-langages. En langue, ... *in many cases there may be uncertainty as to whether a particular operator or argument has selectional frequency for the given word or is a rarer co-occurrent which does not affect its meaning* (Harris, 1991, p. 329–330). Il peut y avoir d'ailleurs désaccord entre les locuteurs et évolution dans le temps par ailleurs en ce qui concerne la probabilité pour un mot d'être lié à un autre comme opérateur ou comme opérande (*ibid.*, p. 16–17). En l'absence d'information quantitative sur les collocations, la nature « floue » des sélections en langue rend difficile le repérage des dépendances opérateur/opérande(s). En revanche, comme les restrictions sont très fortes dans les sous-langages, il est possible de mettre au point des procédures reproductibles permettant non seulement de découvrir ces dépendances opérateur/opérande(s), mais également, en agrégeant ces dépendances, d'aboutir à une sorte de « grammaire sémantique » du sous-langage, sous la forme de combinaisons, de *tuples* de classes d'opérateurs et d'opérandes.

3.2 Méthode d'analyse des sous-langages

L'analyse syntaxique d'un corpus et la normalisation des dépendances sous-jacentes facilitent l'induction de ces grammaires sémantiques.

Constitution du corpus

Les résultats fournis dans Harris *et al.* (1989) s'appuient sur une sélection informée d'une vingtaine d'articles scientifiques dans le domaine de l'immunologie écrits entre 1935 et 1966 (Ryckman, 1990, p. 34). Harris s'est appuyé sur des avis d'immunologistes, dont son frère. Sager *et al.* (1987) utilisent des comptes rendus d'hospitalisation.

Notons que ce sont des textes hautement spécialisés qui sont choisis dans chaque domaine (article scientifique ou données techniques). D'autres documents auraient pu être intégrés (textes de vulgarisation, par exemple). Il y a une homogénéité *de facto* dont la contribution aux résultats n'est pas analysée, sauf erreur.

Normalisation des dépendances syntaxiques

Le corpus médical a été *parsé*, c'est-à-dire analysé automatiquement sur le plan syntaxique, grâce à LSP (parseur du projet éponyme *Linguistic String Project*) de Sager – cf. (Sager, 1981), tandis que le corpus d'immunologie a été analysé à la main. Pour réduire la variation en surface et faciliter la mise en évidence des régularités sélectionnelles, un certain nombre de transformations ont été opérées (passage des nominalisations au verbe sous-jacent, passage du passif à l'actif, etc.). Ces transformations ont été effectuées sous contrôle de spécialistes du domaine (Ryckman, 1990, p. 34). Daladier (1990) insiste en effet sur le caractère incontournable du recours à un expert : les acceptabilités et les transformations dépendent du domaine et la seule compétence de « locuteur natif » n'est pas toujours suffisante. C'est une conception restrictive des transformations qui sont *exclusivement conçues comme des relations entre phrases, et non comme des relations entre structures de constituants (purement formelles) sous-jacentes* (Ryckman, 1990, p. 29) : elles sont relatives à un contexte discursif donné et c'est dans ce cadre que leur apport sémantique est *sémantiquement nul ou constant* (*ibid.*).

Les classes d'opérandes se déduisent en principe de l'analyse du corpus. Pour (Harris *et al.*, 1989), un lexique pré-défini d'opérandes a été néanmoins demandé à des immunologistes, à titre de facilité, dans la mesure où ces classes sont évidentes pour des connaisseurs du domaine. Cf. (Daladier, 1990, p. 75) ou (Ryckman, 1990, p. 35) : *Les catégories de représentation ont été définies sur la base des propriétés de restriction de sélection de mots ou d'expressions linguistiques en partant des catégories définies comme élémentaires par les chercheurs du domaine*. Le travail initial de Sager reposait sur une analyse distributionnelle manuelle. Sager a montré avec ses collègues (Hirschman *et al.*, 1975) que des techniques de clustering permettaient d'obtenir de telles classes à partir de textes parsés. Dans une étape ultérieure, des lexiques médicaux comme *International Classification of Diseases* et *Systematized Nomenclature of Medicine* ont servi de ressources complémentaires (Sager, 1986 ; London, 1987). On constate donc une interaction subtile dans ces travaux entre l'analyse distributionnelle « pure » du corpus et le recours à des

ressources externes. En d'autres termes, l'acquisition de classes sémantiques en sous-langage part rarement d'une situation de table rase. Pour reprendre les termes de la section 2, c'est un apprentissage partiellement supervisé.

Des régularités des phrases élémentaires aux patrons informationnels

A en croire (Sager, 1986, p. 7), *[i]t was then straightforward for a program to substitute class names for class-member occurrences in the sentence trees and to make a table of the operator-argument tuples, sorted alphabetically by operator or by the sublanguage class of each argument*. L'interprétation des tuples produit ce que Sager appelle des types-noyau du sous-langage (*sublanguage kernel-types*), dans son cas une quarantaine, et ce qu'Harris dénomme des patrons informationnels (*information formulas*). Sager définit (*ibid*., p. 6) un type noyau comme *a sublanguage operator class and its argument classes in terms of about a dozen sublanguage noun classes*.

Le résultat attendu nous semble être plutôt des « grammaires sémantiques » , entendues comme ensemble de phrases élémentaires associant chacune une classe d'opérateur et certaines classes d'opérandes, plutôt que des classes sémantiques en tant que telles. Les classes (d'opérateurs ou d'opérandes) servent avant tout à dégager les patrons informationnels.

3.3 Statut des connaissances sémantiques acquises

Le statut possible de ces grammaires sémantiques de sous-langages peut mener à des contradictions. Elles sont présentées d'un côté comme des sortes de langages pivots unifiant les textes relevant d'un même domaine mais dans différentes langues (anglais, français, etc.). D'un autre côté, l'analyse des sous-langages met au jour des évolutions dans les relations opérateurs/opérandes qui reflètent les changements conceptuels du champ disciplinaire. Le premier point de vue correspond en fait à une analyse en synchronie, tandis que le second s'avère plus juste pour les plus longues périodes et les changements correspondants. Une hypothèse commune réconcilie ces deux angles d'attaque : ... *the dependence relation of words reflects what one might consider dependencies within man's perceivable world ... word meanings and co-occurrence selections express a categorizing of perceptions of the world* (Harris, 1991, p. 347).

Les patrons informationnels comme langage pivot

A. Daladier, dans (Harris *et al.*, 1989), applique la même méthodologie à un ensemble de textes en français qui forment avec les articles en anglais ce que l'on appellerait maintenant des *corpus comparables*, c'est-à-dire un ensemble de documents de langues différentes mais obéissant aux mêmes contraintes de thème, de « genre » , de date, etc. La grammaire du sous-langage, constituée d'une quinzaine de classes de mots et d'une douzaine de patrons informationnels, apparaît comme partagée par l'anglais et le français (Ryckman, 1990, p. 35). L'hypothèse d'Harris est que cette convergence pour deux langues se généralise : ... *for a given subscience, the reports and discussions written in one language satisfy much the same special grammar as do papers in the same field written in other languages. The structure of each science language is found to conform to the information in that science rather than to the grammar of the whole language* (Harris, 1988, p. viii). Selon lui, pour chaque science, il y a un tel « langage formuloïde » (*formulaic language*), c'est-à-dire dont les propriétés se rapprochent des notations mathématiques (Harris, 1991, p. 4). Cette grammaire est une *structural representation of the knowledge and opinions in the field* (*ibid*., p. 20).

Les changements de vraisemblance dans les sélections reflètent les évolutions conceptuelles

Sur la longue durée, le langage ne cesse de changer (Harris, 1988, p. 92) (Ryckman, 1990, p. 37). L'évolution du langage se marque dans les sélections : *The most general factor in the varied and changing meanings of words is simply the constant though small change in likelihood — what words are chosen as operators and arguments of other words, and how frequently they are thus used* (Harris, 1991, p. 327). Pister ces évolutions s'avère plus simple pour les sous-langages : ... *the known change of information*

through time is seen in the change of word subclasses and sentence types in the successive articles of [the] period (*ibid.*, p. 286). Le corpus d'immunologie a d'ailleurs été explicitement construit de manière diachronique (1935–1966) pour ... *voir s'il était possible de donner une représentation formelle, commode à utiliser, de l'information contenue dans les articles de ce domaine et permettant également de localiser et de caractériser les désaccords entre les chercheurs du domaine et plus généralement les changements d'information intervenant au cours du temps* (Ryckman, 1990, p. 34). En contraste avec l'analyse synchronique, concentrée sur l'obtention d'une représentation canonique de formules relevant d'une notation structurée, l'approche diachronique est plus sensible aux dimensions sociales et historiques du sens, qui expliquent en particulier le rôle nécessaire du flou : *In certain situations there is need for imprecision, when one is dealing with unsettled questions and with areas where concepts are not fixed because the operations or relations of the science are not adequately understood* (Harris, 1991, p. 297).

4 Sémantique « grossière » et approche harrissienne : éléments d'appréciation croisée

Il est fructueux de chercher à articuler les travaux menés par Harris, Sager, Daladier, et d'autres dans les années quatre-vingts et les démarches en acquisition sémantique automatique, florissantes depuis une décennie. Les approches se complètent. Leur examen croisé permet de mieux maîtriser les conditions de classement sémantique de mots. L'examen des classements issus des traits fournis par les outils de traitements automatiques effectivement disponibles (section 4.1) rend moins cruciale l'obtention d'une analyse syntaxique complète incluant des transformations. A l'inverse, les conséquences observables de la distinction harrissienne entre langue et sous-langages (section 4.2) conduisent à préciser les conditions d'emploi des méthodes d'acquisition automatique.

4.1 Pauvreté linguistique ne nuit pas forcément mais fait loi

Contextes syntaxiques vs *fenêtres de mots*

Les régularités distributionnelles (dépendancielles) observées dans les phrases élémentaires « normalisées » permettent à Harris et à Sager de découvrir des classes sémantiques d'opérateurs et d'opérandes. Nous avons évoqué (section 2) la possibilité, concurrente, d'utiliser des contextes linguistiquement « pauvres » pour partitionner les mots d'un corpus : les *k* mots à gauche et à droite, par exemple. La qualité des classes obtenues en pâtit-elle ?

Grefenstette (1996) compare précisément les résultats obtenus sur un même corpus avec les deux types de contextes. Dans l'approche syntaxique, les contextes d'un nom sont constitués par les adjectifs, les noms et les verbes avec lesquels il rentre dans une relation de dépendance (en position de gouverneur ou de dépendant). Les relations de dépendance sont fournis par l'analyseur robuste que Grefenstette a développé : *Sextant* (Grefenstette, 1994b). Dans l'approche « pauvre » , les contextes d'un nom sont représentés par tous les noms, tous les adjectifs et tous les verbes dans les dix mots avant ou après, et au sein de la même phrase. La pauvreté est donc relative puisque les mots sont étiquetés et lemmatisés. La mesure de distance est celle de Jaccard (pondérée : le nombre d'occurrences de chaque contexte est pris en compte). Grefenstette utilise comme corpus des phrases de l'encyclopédie *Grolier* contenant un des trente hyponymes du mot *institution* (comme *establishment*, *charity*...) dans le dictionnaire sémantique *WordNet* (cf. infra). Le corpus dépasse les 400 000 mots, soit la taille de 4 romans de taille moyenne. Pour pouvoir comparer les deux approches, Grefenstette prend le thesaurus de *Roget* comme pierre de touche. Pour un mot donné et une approche donnée, il regarde si son plus proche voisin (le mot avec lequel la proximité est la plus forte selon l'indice de Jaccard) relève de la même catégorie dans le thesaurus. Si c'est le cas, c'est un succès, dans le cas contraire, un échec.

Les résultats sont en fait nuancés. Ils sont globalement corrélés avec les gammes de fréquences. Les contextes syntaxiques « écrémés » , réduits aux relations de dépendance, donnent de meilleurs résultats pour les 600 mots les plus fréquents. Inversement, pour les formes moins fréquentes, les contextes pauvres débouchent sur davantage de succès. Cette variation tient en fait au nombre de traits disponibles dans chaque méthode pour partitionner les mots, comme en témoignait déjà le tableau 2.1. Les contextes

syntaxiques sont « maigres » et diminuent donc les éléments de rapprochement entre mots. Seuls les mots très fréquents entrent dans suffisamment de contextes pour que cet élagage ne soit pas fatal. Par contre, les mots moins fréquents nécessitent des contextes plus larges pour disposer d'assez de points de convergence avec d'autres mots.

L'expérience de Grefenstette conduit à penser qu'il n'est pas toujours nécessaire de recourir à une analyse syntaxique automatique pour obtenir des partitionnements satisfaisants. La figure 1 va d'ailleurs en ce sens : les regroupements sont obtenus à partir d'une fenêtre limitée aux 5 mots pleins étiquetés et lemmatisés à gauche et à droite de chacun des 50 adjectifs les plus fréquents du corpus examiné. Ce constat peut paraître paradoxal : il n'est pas forcément nécessaire de faire appel à des connaissances linguistiques très élaborées. Il est cependant rassurant : il permet de ne pas exiger de l'analyse syntaxique automatique robuste actuelle plus qu'elle n'est en mesure de donner.

Parsage robuste : des dépendances nombreuses, mais non normalisées

Les données textuelles électroniques réelles sont souvent peu « lissées » (fautes de frappe, de grammaire, variations de registre), comme le montrent ces extrait du même corpus : *essaie d'attrapper tout ce qui est à porter de ses mains et tire ++ dessus malgé sa petittesse – je prefère lui privilegiér son espace vital au Sein de l'incu*[bateur]. Pour traiter les volumes de données actuels dans leur état brut, ont été développées dans les dix dernières années des techniques d'analyse syntaxique robuste (Grefenstette, 1994b ; Vergne, 2000). Elles ne visent pas forcément à construire un arbre syntaxique complet de la phrase. Elles doivent pouvoir fournir des résultats partiels (des fragments d'arbre, ou des relations de dépendance, par exemple) sans se bloquer à cause des scories et des malformations. Par contre, peu de progrès a dans l'immédiat été accompli dans la normalisation de ces dépendances, comme le passage d'une phrase passive à la phrase active correspondante, même si certains travaux mettent en oeuvre des règles de réécriture sur des points précis (Jacquemin, 2001).

Harris faisait appel à des transformations pour ramener les énoncés à des constructions minimales et normalisées opérateur-opérandes. Les parseurs robustes ne redressent pas en ce sens les résultats qu'ils fournissent. Les perturbations qui peuvent en résulter pour la qualité des classements obtenus dépend en fait de la taille du corpus. Sur des corpus de faible taille, de quelques dizaines de milliers de mots, du type de ceux analysés par Harris, la normalisation est probablement cruciale pour augmenter les convergences observées entre les mots. Sur les corpus de plus en plus volumineux qui sont disponibles, il semble en fait que les analyseurs automatiques robustes actuels produisent suffisamment de relations de dépendance pour permettre des rapprochements relativement fiables entre mots.

4.2 Langue et sous-langages : à partir de quelles données langagières classer les mots ?

De nouveaux corpus pour éprouver la distinction langue/sous-langages

Depuis le corpus de Brown University (Francis, 1992), se sont développés, surtout pour l'anglais, des *corpus de référence*. Leur objectif est de fournir un ensemble suffisamment volumineux de données textuelles aux domaines ainsi qu'aux conditions de production et de réception précisément définis, et pouvant « représenter » une grande variété de situations de communication. Au million de « mots » du corpus de Brown ont fait place maintenant les « méga-corpus » (Kennedy, 1998), comme le *British National Corpus* : 100 millions de mots étiquetés, dont 10 millions d'oral transcrit, et pour l'écrit des textes de fiction à partir de 1960 et des textes « informatifs » à partir de 1975. Leur taille en fait des « réservoirs à corpus » : on peut assembler « à façon » les documents destinés à étudier une facette particulière du langage (oral spontané / oral lié à des situations de communication répertoriées, comme le cours, l'entretien d'embauche, l'achat d'un billet de train, etc.).

Dans le même temps, il est désormais aisé de rassembler et de traiter des corpus électroniques spécialisés correspondant à une collectivité donnée, à partir des versions électroniques de publications (journaux, thèses, articles scientifiques) et des documents accessibles sur les sites Web.

Les corpus de référence comme les corpus spécialisées permettent de tester avec un large éventail de données les propositions de Harris sur les restrictions de sélection ou sur l'opposition entre langue et sous-

langages. Ils rendent possible l'examen de la variation de la sélection des opérateurs en fonction des opérandes et vice-versa selon le type d'énoncé (écrit/oral, domaine...). Les travaux actuels sur corpus posent de fait une double question. Est-il vraiment possible d'aboutir à des probabilités de dépendance « en langue » , comme semble le proposer Harris ? Au sein d'un domaine donné, tous les textes disponibles sont-ils pertinents au même titre pour découvrir les classes sémantiques du sous-langage correspondant ?

Les propensions « en langue » existent-elles ?

Biber utilise les divisions d'un corpus de référence en domaines pour montrer que la probabilité d'apparition d'une catégorie morpho-syntaxique donnée est fonction du domaine (Biber, 1993, p. 223). Il indique en outre que les séquences de probabilités de catégories morpho-syntaxiques (bigrammes), tout comme les collocations, varient également avec le domaine (*ibid*., p. 225). Sekine (1998) apporte des faits plaidant dans le même sens. Il examine les performances d'un parseur probabiliste sur 8 domaines du corpus de Brown, selon la relation entre les domaines utilisés pour l'apprentissage et ceux servant au test. Outre la distinction en 8 domaines, il utilise aussi une partition plus grossière en deux classes : fiction / non-fiction. Les performances décroissent dans l'ordre suivant : identité du domaine d'apprentissage et du domaine de test ; appartenance de l'ensemble d'apprentissage et de l'ensemble de test à la même classe grossière ; mélange pour l'ensemble de test et pour l'ensemble d'apprentissage d'extraits de tous les domaines. Le pire résultat correspond à l'apprentissage sur une classe grossière et le test sur l'autre classe.

Le fait que le mélange pour l'ensemble de test et pour l'ensemble d'apprentissage d'extraits de tous les domaines donne de mauvais résultats dans les expériences de Sekine jette au moins un doute sur la possibilité à laquelle semble parfois croire Harris de déterminer pour la langue des probabilités d'emploi d'un mot dans une certaine configuration (avec certaines classes d'opérateurs et d'opérandes). Le fait qu'en reconnaissance automatique de la parole (Jurafsky & Martin, 2000), on ne dispose pas non plus de « modèles de langage » tout terrain, c'est-à-dire peu sensibles au changement de locuteur, de thématique ou de conditions d'interaction orale, contribue aussi à rendre dubitatif sur l'existence de propensions « en langue » .

La variation de registres au sein d'un domaine influe sur les classements sémantiques

L'homogénéité des sous-langages postulée par Harris est elle aussi sujette à caution. L'existence au sein d'un domaine donné de « styles » ou « genres » distincts entraîne vraisemblablement des variations dans l'association des mots entre eux, dans leurs relations de sélection. Par exemple, pour la campagne GRACE d'évaluation d'étiqueteurs morphosyntaxiques pour le français (Paroubek & Rajman, 2000 ; Adda *et al.*, 1999), le corpus comprenait des articles de journaux et des textes littéraires. L'opposition au sein de ces derniers entre les extraits de mémoires (récits de vie) et les extraits de romans, relevant donc du même domaine mais de styles distincts, entraînait des variations de performances des étiqueteurs (Illouz, 1999). On peut faire l'hypothèse, en particulier à la lumière des expériences menées depuis une dizaine d'années en acquisition terminologique, qu'au sein d'un domaine scientifique ou technique, certains types de textes sont plus appropriés que d'autres à la démarche que propose Harris. Nous avons ainsi personnellement constaté, dans un corpus centré sur les maladies coronariennes, *Menelas* (Zweigenbaum, 1994), qu'un extrait de manuel de médecine sur le domaine et des comptes rendus d'hospitalisation étaient plus propices à l'acquisition semi-automatique de classes sémantiques que des lettres de médecins hospitaliers aux médecins traitants des malades en cause, malgré le partage du vocabulaire et d'une partie des constructions.

5 Des classements sémantiques... à dégrossir

La mise en regard de ces deux approches distributionnelles de classement sémantique fait naître de nouvelles perspectives.

5.1 Affiner la constitution des données nécessaires aux classements

La masse grandissante des données textuelles électroniques ne doit pas dissimuler leur hétérogénéité (Illouz *et al.*, 1999). Elle rend crucial le contrôle de cette dernière. L'opposition langue générale / sous-langage qu'ont explorée Harris et ses collaborateurs renvoie à une des dimensions de cette hétérogénéité. D'autres dimensions demandent cependant à être prises en compte, comme les genres ou registres. Un variationisme renouvelé doit s'attacher à explorer ces dimensions, et à caractériser leur influence en acquisition sémantique automatique. Il peut s'appuyer à la fois sur la multiplicité de données langagières électroniques, aux conditions de production bien caractérisées, et sur l'avancée des outils et méthodes de typologie de textes (Biber, 1995 ; Karlgren, 2000).

5.2 Acquisition automatique : des propositions à valider

Comme en témoigne la figure 1, le partitionnement de mots en fonction des proximités distributionnelles débouche sur des regroupements qui nécessitent sans conteste possible un travail humain en aval : émondage et interprétation. Certaines « classes » constituent de simples artefacts à éliminer. Des intrus doivent être enlevés de certains groupes, par ailleurs cohérents. Un regroupement peut rassembler des mots relevant de relations sémantiques hétérogènes (méronymie et synonymie/antonymie, par exemple).

5.3 Des dictionnaires sémantiques électroniques pour le français ?

La lexicographie anglo-saxonne maintient une tradition thesaurique (Miller, 1996) curieusement peu représentée pour le français. Le responsable d'un des rares thesaurus du français note d'ailleurs dans la préface de cet ouvrage : *L'on s'étonnera peut-être que le projet de Roget de prendre la langue dans un réseau conceptuel maillé couvrant méthodiquement l'ensemble des champs notionnels possibles n'ait pas trouvé jusqu'aujourd'hui son équivalent en français. Sans doute y aurait-il là, pour l'historien des mentalités, un motif d'interrogations. Pourquoi un tel projet a-t-il été conçu d'abord dans l'environnement anglo-saxon ? Pure contingence historique ? Effet d'on ne sait quel surmoi cartésien censurant d'emblée un projet de nature essentiellement pragmatique ?* (Péchoin, 1992).

Pour l'anglais, sont au contraire dans le domaine public, sous forme électronique, de nombreux dictionnaires sémantiques, généraux ou spécialisés. En 1985, l'équipe de psycholinguistes autour de Georges Miller a commencé à developper un dictionnaire électronique, *WordNet*. Ce dictionnaire comprend aujourd'hui plus de 170 000 vedettes simples ou « en plusieurs mots » , liées par plus de 350 000 relations, parmi lesquelles synonymie et hyponymie tiennent une place majeure (Fellbaum, 1998 ; Habert, 1998 ; Fellbaum, 1999). *WordNet* (http://www.cogsci.princeton.edu/ wn/) a joué et continuer de jouer un rôle central dans le développement des méthodes d'acquisition sémantique automatique. Terminologies et thesaurus ont également été développés dans de nombreux domaines. Ainsi, la *National Library of Medecine* américaine a lancé en 1986 un projet de mise en relation de diverses terminologies médicales : UMLS (*Unified Medical Language System* – http://www.nlm.nih.gov/research/umls/). Aujourd'hui UMLS regroupe plus de soixante terminologies, totalisant 800 000 concepts et près de deux millions de termes. Chaque concept reçoit une ou plusieurs étiquettes sémantiques issu d'un réseau de 134 types sémantiques.

On imagine l'intérêt de telles ressources en linguistique française. Un exemple entre cent. M. Plénat a dégagé certaines des contraintes phonologiques qui régissent l'emploi du suffixe *-esque* en français (Plénat, 1997). Des hypothèses relativement divergentes ont été proposées sur le sens figuré ou non de ce suffixe (Corbin *et al.*, 1993 ; Bartning & Noailly, 1995). Pour aller plus avant, on aimerait connaître les classes sémantiques privilégiées des noms bases, mais aussi des noms recteurs des dérivés en *-esque*. Encore faut-il pouvoir s'appuyer sur un dictionnaire sémantique de noms.

Il reste du grain à moudre en sémantique à gros grain. Le grain sera d'autant plus fin que s'approfondira le dialogue entre les approches linguistiques et les approches « machinales » . C'est à ce dialogue que nous avons essayé de contribuer.

6 Références

ADDA, G., MARIANI, J., PAROUBEK, P., RAJMAN, M. & LECOMTE, J. (1999). L'action GRACE d'évaluation de l'assignation des parties du discours pour le français. *Langues*, **2**(2), 119–129.

BARTNING, I. & NOAILLY, M. (1995). Pourquoi *-esque* ? *Cahiers de Grammaire*, (20), 87–100.

BIBER, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics*, **19**(2), 243–258.

BIBER, D. (1995). *Dimensions of register variation : a cross-linguistic comparison*. Cambridge: Cambridge University Press.

BORILLO, A. (1996). Exploration automatisée de textes de spécialité : repérage et identification de la relation lexicale d'hyperonymie. *LINX*, (34–35), 113–124.

BRILL, E. & MARCUS, M. (1992). Automatically acquiring phrase structure using distributional analysis. In *Proceedings of the DARPA Speech and Natural Language Workshop*.

CHURCH, K., GALE, W., HANKS, P. & HINDLE, D. (1991). Parsing, word associations and typical predicate-argument relations. In M. TOMITA, Ed., *Current issues in parsing technology*, pp. 103–112. Kluwer Academic Publishers.

CHURCH, K. W. & HANKS, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**(1), 22–29.

CORBIN, D., DAL, G., MÉLIS-PUCHULU, A. & TEMPLE, M. (1993). D'où viennent les sens *a priori* figurés des mots construits ? variations sur *lunette(s)*, *ébéniste* et les adjectifs en *-esque*. *Verbum*, (1–2-3), 65–100.

DACHELET, R. (1994). *Sur la notion de sous-langage*. Thèse de doctorat en sciences du langage, Université Paris VIII, Saint-Denis.

DALADIER, A. (1990). Aspects constructifs des grammaires de Harris. *Langages*, (99), 57–84. A. Daladier (resp.).

C. FELLBAUM, Ed. (1998). *WordNet: an electronic lexical database*. Language, Speech and Communication. Cambridge, Massachusetts: The MIT Press.

FELLBAUM, C. (1999). La représentation des verbes dans le réseau sémantique WordNet. *Langages*, (136), 27–40. Sémantique lexicale et grammaticale – Yvette Yannick Mathieu (ed.).

FIRTH, J. (1957). A synopsis of linguistic theory 1930–1955. *Studies in Linguistic Analysis*, pp. 82–95. Réédité, *Selected Papers of J. R. Firth*, F. Palmer (edt), Longman.

FRANCIS, W. N. (1992). Language corpora B. C. In J. SVARTVIK, Ed., *Directions in Corpus Linguistics*, number 65 in Trends in Linguistics, pp. 17–32. Berlin: Mouton de Gruyter.

GREFENSTETTE, G. (1994a). Corpus-derived first, second and third order affinities. In *EURALEX*, Amsterdam.

GREFENSTETTE, G. (1994b). *Explorations in Automatic Thesaurus Discovery*. Dordrecht, The Netherlands: Kluwer Academic Publisher.

GREFENSTETTE, G. (1996). Evaluation techniques for automatic semantic extraction : Comparing syntactic and window based approaches. In B. BOGURAEV & J. PUSTEJOVSKY, Eds., *Corpus Processing for Lexical Acquisition*, Language, Speech and Communication, chapitre 11, pp. 205–216. Cambridge, Massachusetts: The MIT Press.

HABERT, B. (1998). Compte-rendu de Christiane Fellbaum (resp.) *WordNet, an electronic lexical database*, The MIT Press, 1998. *TAL*, **39**(2), 152–155.

HABERT, B. & ZWEIGENBAUM, P. (2002). Contextual acquisition of information categories: what has been done and what can be done automatically? In B. NEVIN, Ed., *The Legacy of Zellig Harris: Language and information into the 21st century*, volume 2. Computability of language and computer applications. Amsterdam: John Benjamins. A paraître.

HARRIS, Z. (1988). *Language and information*. New York: Columbia University Press.

HARRIS, Z., GOTTFRIED, M., RYCKMAN, T., MATTICK JR, P., DALADIER, A., HARRIS, T. & HARRIS, S. (1989). *The Form of Information in Science, Analysis of Immunology Sublanguage*, volume 104 of *Boston Studies in the Philosophy of Science*. Dordrecht, The Netherlands: Kluwer Academic Publisher.

HARRIS, Z. S. (1991). *A theory of language and information. A mathematical approach*. Oxford: Oxford University Press.

HEARST, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Actes, 14th International Conference on Computational Linguistics (COLING'92)*, pp. 539–545, Nantes.

HIRSCHMAN, L., GRISHMAN, R. & SAGER, N. (1975). Grammatically-based automatic word class formation. *Information Processing & Management*, **11**(1/2), 39–57.

IDE, N. & VÉRONIS, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, **24**(1), 1–40.

ILLOUZ, G. (1999). Méta-étiqueteur adaptatif : vers une utilisation pragmatique des ressources linguistiques. In P. AMSILI, Ed., *Actes de TALN'99 (Traitement Automatique des Langues Naturelles)*, pp. 185–194, Cargèse: ATALA.

ILLOUZ, G., HABERT, B., FLEURY, S., FOLCH, H., HEIDEN, S. & LAFON, P. (1999). Maîtriser les déluges de données hétérogènes. In A. CONDAMINES, C. FABRE & M.-P. PÉRY-WOODLEY, Eds., *Corpus et traitement automatique des langues : pour une réflexion méthodologique*, pp. 37–46, Cargèse: ATALA.

JACQUEMIN, C. (2001). *Spotting and discovering terms through natural language processing*. Cambridge, Massachusetts: The MIT Press.

JURAFSKY, D. & MARTIN, J. H. (2000). *Speech and language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Artificial Intelligence. Upper Saddle River, New Jersey: Prentice Hall.

KARLGREN, J. (2000). *Stylistic Experiments for Information Retrieval*. Phd in computational linguistics, Swedish Institute of Computer Science, Stockolm, Sweden.

KENNEDY, G. (1998). *An introduction to corpus linguistics*. Studies in language and linguistics. London: Longman.

LEBART, L., MORINEAU, A. & PIRON, M. (1997). *Statistique exploratoire multidimensionnelle*. 2e cycle. Paris: Dunod, 2ème edition.

LEEMAN, D. (1996). Le « sens » et l'« information » chez Harris. *LINX*, pp. 209–220.

LONDON, J. (1987). The healthcare lexicon. In N. SAGER, C. FRIEDMAN & M. S. LYMAN, Eds., *Medical Language Processing : Computer Management of Narrative Data*, chapter 6, pp.137–144. Addison-Wesley.

LOSEE, R. M. (1998). *Text Retrieval and Filtering: Analytic Models of Performance*. Information Retrieval. Dordrecht: Kluwer Academic Publishers.

MANNING, C. D. & SCHÜTZE, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

MILLER, G. A. (1996). *The Science of Words*. New York: Scientific American Library.

MITCHELL, T. M. (1997). *Machine Learning*. Computer Science. New York: McGraw-Hill.

MORIN, E. (1999). *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Doctorat en informatique, Institut de Recherche en Informatique de Nantes, Nantes.

NAZARENKO, A., ZWEIGENBAUM, P., HABERT, B. & BOUAUD, J. (2001). Corpus-based extension of a terminological semantic lexicon. In D. BOURIGAULT, C. JACQUEMIN & M.-C. L'HOMME, Eds., *Recent Advances in Computational Terminology*, Natural Language Processing, chapter 16, pp. 327–351. Amsterdam: John Benjamins.

NEVIN, B. (1993). A minimalist program for linguistics: The work of Zellig Harris on meaning and information. *Historiographia Linguistica*, **XX**(2/3), 355–398.

PAROUBEK, P. & RAJMAN, M. (2000). étiquetage morpho-syntaxique. In J.-M. PIERREL, Ed., *Ingéniérie des langues*, Informatique et systèmes d'information, chapter 5, pp. 131–150. Paris: Hermès Science.

D. PÉCHOIN, Ed. (1992). *Thésaurus Larousse – des idées aux mots, des mots aux idées*. Paris: Larousse, 2ème edition.

PEREIRA, F. (2000). Formal grammar and information theory: together again? *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, (358), 1239–1253. Royal Society, London.

PLÉNAT, M. (1997). Analyse morpho-phonologique d'un corpus d'adjectifs dérivés en *-esque*. *French Language Studies*, (7), 163–179.

RYCKMAN, T. (1990). De la structure d'une langue aux structures de l'information dans le discours et dans les sous-langages scientifiques. *Langages*, (99), 21–28. A. Daladier (resp.).

SAGER, N. (1981). *Natural Language Information Processing : A Computer Grammar of English and Its Applications*. Addison Wesley.

SAGER, N. (1986). Sublanguage : Linguistic phenomenon, computational tool. In R. GRISHMAN & R. KITTREDGE, Eds., *Analyzing Language in Restricted Domains : Sublanguage Description and Processing*, chapter 1, pp. 1–18. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

SAGER, N., FRIEDMAN, C. & (EDITORS), M. S. L. (1987). *Medical Language Processing : Computer Management of Narrative Data*. Addison-Wesley.

SEKINE, S. (1998). The domain dependence of parsing. In *Fifth Conference on Applied Natural Language Processing*, pp. 96–102, Washington: Association for Computational Linguistics.

SOWA, J. F. (2000). *Knowledge Representation. Logical, Philosophical and Computational Foundations.* Pacific Grove, CA: Brooks/Cole.

VERGNE, J. (2000). *Trends in Robust Parsing*. GREYC – Université de Caen, Nancy. COLING 2000 Tutorial.

ZWEIGENBAUM, P. (1994). MENELAS: an access system for medical records using natural language. *Computer Methods and Programs in Biomedicine*, **45**, 117–120.