dans l'échantillon, bien qu'ils soient donnés par Pollard et Sag comme non acceptables ([*to regard* NP NP] ; et [*to regard* NP VP[inf]] : *Conservatives argue that the Bible regards homosexuality to be a sin*)[18]. L'analyse de Manning souligne deux points. D'abord, « … la frontière de la grammaticalité a, semble-t-il, été tracée à un point de basse fréquence assez arbitraire. Grossièrement, ce qui arrive plus d'une fois sur 100 est considéré comme grammatical, tandis que ce qui arrive moins fréquemment est jugé non grammatical. » Ensuite, à l'inverse, Pollard et Sag mettent sur le même plan les différents cadres de sous-catégorisation jugés acceptables alors que [*to regard* NP as NP] (*We regard Kim as an acceptable candidate*) figure dans plus de 75 % des phrases de l'échantillon tandis que la fréquence des autres est beaucoup plus faible ([*to regard* NP as AdjP] : ≈ 17 % ; [*to regard* NP as VP[ing]] : ≈ 4 % ; [*to regard* NP as PP] : ≈ 2 %)[19].

3. Le dernier Harris : au carrefour de la linguistique formelle et de la théorie de l'information

L'utilisation depuis une dizaine d'années de méthodes globalement distributionnelles en acquisition sémantique automatique (Grefenstette, 1994) s'est accompagnée d'un « retour à Harris ». Ce retour à Harris privilégie, semble-t-il, le distributionnalisme des années 1950-1960[20]. On peut, sinon à rebours, du moins de manière complémentaire, souligner, à la suite de Pereira (2000, p. 1240), qu'Harris, en particulier dans ses derniers travaux (Harris, 1988 ; Harris *et al.*, 1989 ; Harris, 1991), « a développé ce qui est probablement la proposition la plus achevée pour un mariage de la linguistique avec la théorie de l'information ». C'est cet apport spécifique que nous rappelons[21]. Nous nous appuyons sur la section 2 de (Habert & Zweigenbaum, 2002)[22] pour la présentation des positions de Harris sous cet angle. 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).

Nous présentons dans le paragraphe 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

18. On retrouve l'analyse d'Auroux (1998, p. 197) : « La règle est une hypothèse sur les faits, les faits contiennent aussi bien du possible que de l'impossible. »

19. Ces proportions sont estimées par nous à partir de la figure fournie par Manning.

20. Par exemple, Church *et al.* (1991a) renvoient à (Harris, 1968) et Brill & Marcus (1992) à (Harris, 1951). Étonnante *a contrario* l'absence totale de Harris dans la bibliographie de (Resnik, 1993), étude probabiliste marquante sur les restrictions de sélection.

21. (Daladier, 1990), (Ryckman, 1990) et (Dachelet, 1994) constituent les principales introductions en français.

22. L'essentiel de l'article cité est centré sur une évaluation des méthodes proposées par Harris en matière d'acquisition de catégories sémantiques au regard de l'état de l'art en traitement automatique du langage.

la normalisation des relations syntaxiques sous-jacentes permettent de dégager des catégories et des patrons sémantiques (paragraphe 3.2). Ces catégories et ces patrons reflètent le monde perçu, son évolution et les relations qui l'organisent (paragraphe 3.3).

3.1. *Hypothèses fondamentales*

3.1.1. *Le sens : un résultat, pas un point de départ*

Pour 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 : 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. Plus largement, Harris refuse la réduction du langage à un *code*, défini comme *une bijection entre des expressions déjà bien formées (le « message ») et les éléments du chiffrage choisi* (Ryckman, 1990, p. 25)[23]. Pour lui, il n'existe pas, pour analyser le langage, de métalangue externe à la langue. Si bien que *pour 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 »* (*ibid.*).

Cette position ne conduit pas Harris à abandonner pour autant toute investigation de type sémantique[24]. 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)*[25] (*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).

---

23. Cf. aussi Nevin (1993, p. 358-359).

24. M. Gross, qui a prolongé à sa façon en France la démarche de Harris (Gross, 1996), est plus réservé sur ce point (Gross, 1981).

25. Cf. Parlebas (1992).

On pourrait vouloir rapprocher cette position de la citation de Firth (1957) sur les collocations – *you shall know a word by the company it keeps*[26], rendue en particulier fameuse par l'utilisation de l'information mutuelle pour mesurer l'attirance entre mots (Church *et al.* 1991a, 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)[27]. En fait, 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).

3.1.2. *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* ou de *propension* (*likelihood*) de (Harris, 1988), rappelée dans (Pereira, 2000, p. 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*. » C'est cette relation entre vraisemblance d'associations opérateur-opérande(s) et information qui rapproche la démarche harrissienne de la théorie de l'information[28] : *[d]ans la mesure où il n'y a pas de métalangue disponible pour la description grammaticale, toute restriction sur les combinaisons contenues dans la grammaire doit correspondre à ou être corrélée avec une différence d'information, une différence reconnue par les locuteurs de la langue* (Ryckman, 1990, p. 27-28). Ou encore : *In the absence of an external metalanguage, the entities of each language can be identified only if the sounds, markers, or words of which they are composed do not occur randomly in utterances of the language. That is, the entities can be recognized only if not all combinations occur, or are equally probable. This condition is indeed satisfied by languages. A necessary*

26. Cf. Stubbs (1996) pour une présentation de la tradition firthéenne et des travaux sur les collocations.

27. On notera d'ailleurs que pour Church *et al.* (1991b, p. 159), le Harris de l'analyse distributionnelle est présenté surtout comme participant de l'« air du temps firthéen ».

28. Même si Harris (1991) ne semble pas citer Shannon et si Ryckman (1990, p. 21-27), sur le rapport entre les deux approches, met avant tout l'accent sur le refus, chez Harris, de réduire le langage à un code.

step, then, towards understanding language structure is to distinguish the combinations of elements that occur in the utterances of a language from those that do not : that is, to characterize their departures from randomness (Harris, 1988, p. 3). 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).

3.1.3. *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 (1986, p. 2), 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*. 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 (*language 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 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.

3.2.1. *Constitution du corpus*

Les résultats fournis dans (Harris *et al.*, 1989) s'appuient sur une sélection informée[29] d'une vingtaine d'articles scientifiques dans le domaine de l'immunologie écrits

29. Harris s'est appuyé sur des avis d'immunologistes, dont son frère, qui a contribué avec S. Harris à un état de la question scientifique *The cellular source of antibody : a review* (*ibid.*, p. 192-215).

entre 1935 et 1966 (Ryckman, 1990, p. 34). Les contributions de Sager *et al.* (1987) reposent sur 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.

3.2.2. *Normalisation des dépendances syntaxiques*

Le corpus médical a été parsé, grâce à LSP (parseur du projet éponyme *Linguistic String Project*) de Sager (1981), tandis que le corpus d'immunologie a été analysé à la main[30]. Pour réduire la variation en surface et faciliter la mise en évidence des régularités sélectionnelles, un certain nombre de transformations ont été opérées[31] (passage des nominalisations au verbe sous-jacent, passage du passif à l'actif, etc.).

Les classes d'opérandes se déduisent en principe de l'analyse du corpus. Dans (Harris *et al.*, 1989), un lexique prédéfini d'opérandes a été néanmoins demandé à des immunologistes (Daladier, 1990, p. 75)[32], par simple commodité, dans la mesure où ces classes sont évidentes pour des connaisseurs du domaine. Le travail initial de Sager et de ses collègues reposait sur une analyse distributionnelle manuelle. Ils ont montré dans (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, p. 6) (London, 1987). On constate donc une interaction subtile dans ces travaux entre l'analyse distributionnelle du corpus et le recours à des ressources externes. En d'autres termes, l'acquisition de classes sémantiques dans un sous-langage part rarement d'une situation de table rase.

3.2.3. *Des régularités des phrases élémentaires aux patrons informationnels*

A en croire Sager (1986, p. 7), *it was then straightforward for a program to substitute class names for class-member occurrences in the sentence trees and to make*

---

30. On comprend dans ces conditions la taille restreinte de ces corpus au regard de nos critères actuels.

31. Éventuellement sous contrôle de spécialistes du domaine (Ryckman, 1990, p. 34). Daladier (1990) insiste 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.*).

32. Cf. 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.*

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'opérateurs 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).

3.3.1. *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[33]. 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,

33. Ensemble de documents de langues différentes obéissant aux mêmes contraintes de thème, de « genre », de date, etc.

1991, p. 4). Cette grammaire est une *structural representation of the knowledge and opinions in the field* (*ibid*., p. 20).

3.3.2. *Les changements de vraisemblance 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 pouvoir rendre manifestes les désaccords et les changements dans un champ[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* (*ibid.*, p. 297).

4. Régler les règles : associer règles « catégoriques »/« propensions » dans des « espaces de règles »

Il y a une dizaine d'années, Harris dessinait un programme fondant les grammaires sur les propensions (*likelihood*) de dépendances des mots entre eux en termes d'opérateurs/opérandes. Il s'éloignait en cela d'une vision purement « catégorique » des règles linguistiques (en termes d'acceptabilité/inacceptabilité). Il ancrait ces grammaires dans les sélections observées plus que dans l'intuition du locuteur natif.

Ce cadre théorique retrouve aujourd'hui une nouvelle fécondité. Il est en effet en harmonie avec les travaux des dix dernières années sur l'acquisition de restrictions de sélection et de cadres de sous-catégorisation (Manning & Schütze, 1999, ch. 8). Ainsi Resnik (1993, ch. 4) corrobore-t-il de manière empirique, sur des bases probabilistes, les hypothèses de Harris sur l'effacement : plus fortes sont les contraintes qu'un verbe place sur son objet, plus ce dernier tend à s'effacer. L'accent mis sur les relations entre *tel* opérateur et *tel(s)* opérande(s) chez Harris trouve un écho naturel dans la focalisa-

34. 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).

partir du moment où l'on ajoute ce que Manning dénomme des « structures cachées » (les variables cachées de Pereira (2000)), ce qui est en cours dans la constitution de corpus annotés sous de multiples angles (arbres syntaxiques (Marcus *et al.*, 1993) ; relations de co-référence ; étiquettes sémantiques (Landes *et al.*, 1998) ; phonétisation, indications prosodiques et métriques (Beaudouin, 2002), etc.), ces corpus enrichis permettent le test d'hypothèses sophistiquées[46] mais aussi permettent de mettre en évidence des phénomènes ou des corrélations inattendues. Ils permettent également de progresser vers de nouvelles grammaires articulant règles catégoriques et propensions observées.

Pour laisser la parole à Pereira (2000, p. 1250) : « ... il est bien possible que nous assistions à l'émergence d'une nouvelle version du programme harrissien, dans lequel des modèles computationnels contraints par des considérations grammaticales définissent des grandes classes [*broad classes*] de grammaires possibles, et des principes empruntés à la théorie de l'information spécifient comment ces modèles s'ajustent aux données linguistiques attestées. »

Merci aux trois relecteurs. Leurs regards et leurs questionnements nous ont notablement permis d'avancer dans la réflexion. La qualité du rendu typographique final tient à la vigilance du prote d'Hermès Science Publications.

5. Bibliographie

ABNEY S., « Partial parsing via finite-state cascades », *Natural Language Engineering*, vol. 2, n° 4, 1996, p. 337-344.

ADDA G., MARIANI J., PAROUBEK P., LECOMTE J., « Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morphosyntaxiques pour le français », AMSILI P., Ed., *Actes de TALN'99 (Traitement Automatique des Langues Naturelles)*, Cargèse, 12-17 juillet 1999, ATALA, p. 15-24.

ADDA G., MARIANI J., PAROUBEK P., RAJMAN M., LECOMTE J., « L'action GRACE d'évaluation de l'assignation des parties du discours pour le français », *Langues*, vol. 2, n° 2, 1999, p. 119-129.

AUROUX S., *La raison, le langage et les normes*, Presses Universitaires de France, 1998.

BAAYEN R. H., *Word Frequency Distribution*, Kluwer Academic Publishers, 2001.

BARKEMA H., « Idiomaticy in English NPs », AARTS J., DE HAAN P., OOSTDIJK N., Eds., *English language corpora : design, analysis and exploitation*, p. 257-278, Rodopi, Amsterdam, 1993.

BARKEMA H., « Determining the syntactic flexibility of idioms », FRIES U., TOTTIE G., SCHNEIDER P., Eds., *Creating and using English language corpora*, p. 39-52, Rodopi, Amsterdam, 1994.

46. Manning teste effectivement les hypothèses de Pollard & Sag (1994) sur les cadres de sous-catégorisation de *to regard*.

BEAUDOUIN V., *Mètre et rythmes du vers classique : Corneille et Racine*, Honoré Champion, 2002.

BIBER D., « Using register-diversified corpora for general language studies », *Computational Linguistics*, vol. 19, n° 2, 1993, p. 243-258.

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

BRILL E., MARCUS M., « Automatically Acquiring Phrase Structure Using Distributional Analysis », *Proceedings of the DARPA Speech and Natural Language Workshop*, 1992.

CHIERCHIA G., MCCONNEL-GINET S., *Meaning and Grammar*, The MIT Press, Cambridge, Massachusetts, 1990.

CHOMSKY N., *Syntactic Structures*, Mouton, La Haye, 1957.

CHURCH K. W., HANKS P., « Word Association Norms, Mutual Information, and Lexicography », *Computational Linguistics*, vol. 16, n° 1, 1990, p. 22-29.

CHURCH K., GALE W., HANKS P., HINDLE D., « Parsing, word associations and typical predicate-argument relations », TOMITA M., Ed., *Current issues in parsing technology*, p. 103-112, Kluwer Academic Publishers, Dordrecht, 1991.

CHURCH K., GALE W., HANKS P., HINDLE D., « Using Statistics in Lexical Analysis », ZERNIK U., Ed., *Lexical Acquisition*, p. 115-164, Lawrence Erlbaum Ass., Hillsdale, NJ, 1991.

CORBIN P., « De la production des données en linguistique introspective », DESSAUX-BERTHONEAU A.-M., Ed., *Théories linguistiques et traditions grammaticales*, p. 121-179, Presses Universitaires de Lille, Villeneuve-d'Asq, 1980.

CORBIN D., « Entre les mots possibles et les mots existants : les unités lexicales à faible probabilité d'actualisation », CORBIN D., FRADIN B., HABERT B., KERLEROUX F., PLÉNAT M., Eds., *Mots possibles et mots existants*, Lille, avril 1997, p. 79-90.

CUNNINGHAM H., « A definition and short history of Language Engineering », *Natural Language Engineering*, vol. 5, n° 1, 1999, p. 1-16.

DACHELET R., « Sur la notion de sous-langage », Thèse de doctorat en sciences du langage, Université Paris VIII, Saint-Denis, décembre 1994.

DAGAN I., PEREIRA F., LEE L., « Similarity-Based Estimation of Word Cooccurrences Probabilities », *32nd Annual Meeting*, Nouveau-Mexique, Etats-Unis, 27-30 juin 1994, Association for Computational Linguistics, p. 272-278.

DALADIER A., « Aspects constructifs des grammaires de Harris », *Langages*, n° 99, 1990, p. 57-84, A. Daladier (ed.).

FIRTH J., « A synopsis of linguistic theory 1930-1955 », *Studies in Linguistic Analysis*, 1957, p. 82-95, Philological Society, Réédité, *Selected Papers of J. R. Firth*, F. Palmer (ed.), Longman.

FRANCIS W. N., « Language Corpora B. C. », SVARTVIK J., Ed., *Directions in Corpus Linguistics*, p. 17-32, Mouton de Gruyter, 1992.

FUCHS C., *Linguistique et traitement automatique des langues*, Hachette, 1993.

GADET F., « Variation et hétérogénéité », *Langages*, n° 108, 1992, p. 5-15, Hétérogénéité et variation : Labov, un bilan, Françoise Gadet (ed.).

GAZDAR G., « Paradigm merger in natural language processing », MILNER R., WAND I., Eds., *Computing tomorrow : Future research directions in computer science*, p. 88-109,

Cambridge University Press, Cambridge, 1996.

GREFENSTETTE G., « Corpus-derived first, second and third order affinities », *EURALEX*, Amsterdam, August 1994.

GROSS M., « Présentation », *in* Jean-Paul Boons, Alain Guillet, Christian Leclère, *La structure des phrases simples en français*, Droz, Genève, 1976.

GROSS M., « Les bases empiriques de la notion de prédicat sémantique », *Langages*, n° 63, 1981, p. 7-52.

GROSS M., « Lexicon Grammar », BROWN K., MILLER J., Eds., *Concise Encyclopedia of syntactic theories*, p. 244-258, Pergamon, Cambridge, 1996.

HABERT B., ZWEIGENBAUM P., « Contextual Acquisition of Information Categories : what has been done and what can be done automatically ? », NEVIN B., Ed., *The Legacy of Zellig Harris : Language and information into the 21st century*, vol. 2. Computability of language and computer applications, John Benjamins, Amsterdam, 2002.

HARRIS Z., GOTTFRIED M., RYCKMAN T., MATTICK JR P., DALADIER A., HARRIS T., HARRIS S., *The Form of Information in Science, Analysis of Immunology Sublanguage*, Kluwer Academic Publisher, Dordrecht, Pays-Bas, 1989.

HARRIS Z., *Structural Linguistics*, University of Chicago Press, Chicago, 1951.

HARRIS Z., *Mathematical structures of language*, John Wiley, New York, 1968.

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

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

HIRSCHMAN L., GRISHMAN R., SAGER N., « Grammatically-based Automatic Word Class Formation », *Information Processing & Management*, vol. 11, n° 1/2, 1975, p. 39-57.

ILLOUZ G., HABERT B., FLEURY S., FOLCH H., HEIDEN S., LAFON P., « Maîtriser les déluges de données hétérogènes », CONDAMINES A., FABRE C., PÉRY-WOODLEY M.-P., Eds., *Corpus et traitement automatique des langues : pour une réflexion méthodologique*, Cargèse, 12-17 juillet 1999, ATALA, p. 37-46.

ILLOUZ G., « Méta-étiqueteur adaptatif : vers une utilisation pragmatique des ressources linguistiques », AMSILI P., Ed., *Actes de TALN'99 (Traitement Automatique des Langues Naturelles)*, Cargèse, 12-17 juillet 1999, ATALA, p. 185-194.

ILLOUZ G., « Typage de données textuelles et adaptation des traitements linguistiques. Application à l'annotation morpho-syntaxique », Doctorat d'informatique, Université Paris-Sud, Orsay, décembre 2000.

JURAFSKY D., MARTIN J. H., *Speech and language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Prentice Hall, 2000.

KARLGREN J., « Stylistic Experiments for Information Retrieval », PhD in Computational Linguistics, Swedish Institute of Computer Science, Stockolm, Sweden, 2000.

KARTTUNEN L., CHANOD J.-P., GREFENSTETTE G., SCHILLE A., « Regular expressions for language engineering », *Natural Language Engineering*, vol. 2, n° 4, 1996, p. 305-328.

KENNEDY G., *An introduction to corpus linguistics*, Longman, 1998.

LANDES S., LEACOCK C., TENGI R. I., « Building Semantic Concordances », FELLBAUM C., Ed., *WordNet : an electronic lexical database*, p. 199-216, The MIT Press, 1998.

LEEMAN D., « Le “sens” et l’“information” chez Harris », *LINX*, 1996, p. 209-220, « Du dire et du discours ». Hommage à Denise Maldidier.

LONDON J., « The Healthcare Lexicon », SAGER N., FRIEDMAN C., LYMAN M. S., Eds., *Medical Language Processing : Computer Management of Narrative Data*, Addison-Wesley, 1987.

MAIR C., « Changing patterns of complementation, and concomitant grammaticalisation, of the verb *help* in present-day British English », AARTS B., MEYER C. F., Eds., *The verb in contemporary English. Theory and description*, p. 258-271, Cambridge University Press, Cambridge, 1995.

MANI I., MAYBURY M. T., Eds., *Advances in Automatic Text Summarization*, The MIT Press, Cambridge, Massachusetts, 1999.

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

MANNING C. D., « Probabilistic syntax », BOD, HAY, JANNEDY, Eds., *Probabilistic Linguistics*, The MIT Press, Cambridge, Massachusetts, 2002, A paraître.

MARCUS M., SANTORINI B., MARCINKIEWICZ M. A., « Building a Large Annotated Corpus of English : The Penn Treebank », *Computational Linguistics*, vol. 19, n° 2, 1993, p. 313-330.

MAYNARD D., TABLAN V., CUNINGHAM H., URSU C., SAGGION H., BONTCHEVA K., WILKS Y., « Architectual elements for language engineering robustness », *Natural Language Engineering*, vol. 8, n° 2/3, 2002, p. 257-274.

MILNER J.-C., *Introduction à une science du langage*, Seuil, 1989.

NEVIN B., « A Minimalist Program for Linguistics : The Work of Zellig Harris on Meaning and Information », *Historiographia Linguistica*, vol. XX, n° 2/3, 1993, p. 355-398.

PARLEBAS P., *Sociométrie, réseaux et communications*, Presses Universitaires de France, Paris, 1992.

PAROUBEK P., RAJMAN M., « Étiquetage morpho-syntaxique », PIERREL J.-M., Ed., *Ingéniérie des langues*, p. 131-150, Hermès Science Publications, 2000.

PEREIRA F., « Formal grammar and information theory : together again ? », *Philosophical Transactions : Mathematical, Physical and Engineering Sciences*, n° 358, 2000, p. 1239-1253, Royal Society, London.

PIERREL J.-M., Ed., *Ingéniérie des langues*, Hermès Science Publications, 2000.

POLLARD C., SAG I. A., *Head-driven phrase structure grammar*, The University of Chicago Press, 1994.

RESNIK P., « Selection and Information : A Class-Based Approach to Lexical Relationships », PhD thesis, University of Pennsylvania, December 1993, (Institute for Research in Cognitive Science report IRCS-93-42).

ROCHE E., SCHABES Y., Eds., *Finite-state language processing*, The MIT Press, Cambridge, Massachusetts, 1997.

ROSENFELD R., « Incorporating linguistic structure into statistical language models », *Philosophical Transactions : Mathematical, Physical and Engineering Sciences*, n° 358, 2000, p. 1311-1324, Royal Society, London.

RYCKMAN T., « De la structure d’une langue aux structures de l’information dans le discours et dans les sous-langages scientifiques », *Langages*, n° 99, 1990, p. 21-28, A. Daladier

(ed.).

SAGER N., FRIEDMAN C., LYMAN M. S., Eds., *Medical Language Processing : Computer Management of Narrative Data*, Addison-Wesley, Reading, Massachusetts, 1987.

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

SAGER N., « Sublanguage : Linguistic Phenomenon, Computational Tool », GRISHMAN R., KITTREDGE R., Eds., *Analyzing Language in Restricted Domains : Sublanguage Description and Processing*, Lawrence Erlbaum Associates, 1986.

SAMPSON G., « The role of taxonomy in language engineering », *Philosophical Transactions : Mathematical, Physical and Engineering Sciences*, n° 358, 2000, p. 1339-1355, Royal Society, London.

SEKINE S., « The Domain Dependence of Parsing », *Fifth Conference on Applied Natural Language Processing*, Washington, mars-avril 1998, Association for Computational Linguistics, p. 96-102.

SILBERZTEIN M., *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Masson, 1993.

SPARCK JONES K., GALLIERS J. R., *Evaluating Natural Language Processing Systems. An Analysis and Review*, N° 1083, Springer-Verlag, 1996.

SPARCK JONES K. I. B., GAZDAR G. J. M., NEEDHAM R. M., « Introduction : combining formal theories and statistical data in natural language processing », *Philosophical Transactions : Mathematical, Physical and Engineering Sciences*, n° 358, 2000, p. 1227-1238, Royal Society, London.

SPARCK JONES K., « Automatic language and information processing : rethinking evaluation », *Natural Language Engineering*, vol. 7, n° 1, 2001, p. 29-46.

STUBBS M., *Text and Corpus Analysis*, Blackwell, Oxford, 1996.

ZWEIGENBAUM P., « MENELAS : an Access System for Medical Records using Natural Language », *Computer Methods and Programs in Biomedicine*, vol. 45, 1994, p. 117-120.