BENOÎT HABERT
ADELINE NAZARENKO
ANDRÉ SALEM

Les linguistiques de corpus

REMERCIEMENTS

Ce livre doit beaucoup aux laboratoires dans lesquels nous avons travaillé, l'Équipe de Linguistique et Informatique (ELI) de l'École Normale Supérieure de Fontenay/St Cloud (Equipe d'Accueil 463), le Laboratoire d'Informatique de Paris-Nord (URA 1507 – CNRS et Université Paris 13), l'UPRES SYLED (EA 2290 – Université Sorbonne nouvelle Paris 3) et l'UMR 9952 Lexicométrie et Textes Politiques (CNRS – INaLF et ENS de Fontenay/St Cloud).

Nous remercions particulièrement Christiane Marchello-Nizia (ELI), pour son appui chaleureux.

Merci à ceux qui ont complété notre documentation et notre information : Andrée Borillo, Jacques Bouaud, Anne Daladier, Fernande Dupuis, Marc El-Bèze, Fabrice Issac, Sylvain Kahane, Dominique Labbé, Ludovic Lebart, Monique Lemieux, Elie Naulleau, Jean-Marie Marandin et Jean Véronis.

Merci aussi à Pierrette Habert et à Serge Heiden pour leur soutien technique et leur conseils.

Merci enfin à nos collègues et proches qui nous ont relus avec une générosité vigilante : Sophie Aslanidès, Didier Bourigault, Cécile Fabre, Serge Fleury, Helka Folch, Christian Jacquemin, Lucie Langlois, Isabelle Moulinier, Christiane Marchello-Nizia, Sandrine Oriez, Marie-Paule Péry-Woodley et Pierre Zweigenbaum.

INTRODUCTION

1. LE REGAIN D'INTERET POUR LES CORPUS

De vastes corpus de textes électroniques *étiquetés* (chaque mot est assorti d'une étiquette morpho-syntaxique) et parfois munis d'arbres syntaxiques (on parle alors de corpus *arborés*) sont aujourd'hui disponibles pour l'anglais et pour l'américain. Les outils d'interrogation de ces corpus *enrichis* ainsi que les outils d'annotation proprement dits (étiqueteurs, analyseurs syntaxiques, etc.) se répandent. Depuis quelque temps déjà, on trouve dans le domaine public des étiqueteurs pour l'anglais qui permettent de catégoriser des textes préalablement saisis sur support magnétique (Cutting *et al.*, 1992 ; Brill, 1995). Leurs équivalents pour le français apparaissent.

Ce qui est neuf, ce n'est pas l'utilisation de corpus électroniques. En France, un fonds de quelque 160 millions de mots a ainsi été patiemment constitué à l'Institut National de la Langue Française (INaLF – CNRS) depuis les années soixante et constitue une base textuelle désormais accessible en ligne : *Frantext*. Ce fonds a servi en particulier à la rédaction des dix-sept volumes du *Trésor de la Langue Française*.

La nouveauté réside dans l'enrichissement des corpus, l'accroissement de leur taille et dans l'accessibilité effective des corpus et des outils. D'abord, les corpus ne sont plus des suites de mots « nus », c'est-à-dire de simples chaînes de caractères, mais ils sont *annotés* (ou encore *enrichis*). Nous entendons par là l'ajout d'information, de quelque nature qu'elle soit : morphologique, syntaxique, sémantique, prosodique, critique ... Le niveau d'annotation progresse régulièrement. Les années quatre-vingts ont été consacrées à l'étiquetage morpho-syntaxique. La décennie actuelle voit se développer les corpus arborés. Les annotations sémantiques émergent et vont se répandre. Ensuite, la taille de ces corpus ne cesse de croître. K. Church et R. Mercer (1993) notent à ce propos : « Il y a juste 10 ans, le corpus de ***Brown***[1], avec son million de mots, était considéré comme un grand corpus [...]. Aujourd'hui, de nombreux centres de recherche disposent de données textuelles de millions voire de milliards de mots. » Le ***British National Corpus*** (***BNC***) comprend par exemple 100 millions de mots étiquetés. Enfin, ces ressources sont désormais accessibles aux chercheurs universitaires pour des coûts raisonnables et ne sont plus réservées aux seuls centres de recherche industriels ou aux organismes qui ont constitué et mis au point ces données et ces outils.

[1] Il s'agit de Brown University (USA).

Faut-il voir dans cet engouement actuel pour les corpus le retour aux débuts de la linguistique structurale américaine des années cinquante ? Après l'accent chomskyen sur la formalisation et l'intuition du locuteur natif, la revanche de l'empirisme ? Le découragement serait de mise s'il y avait effectivement piétinement et ressassement. Or, l'étude des origines de ces travaux le montre, ce sont les discontinuités qui l'emportent, ainsi que la diversité, voire l'éclatement, des horizons théoriques et des réalisations pratiques.

2. À QUOI SERVENT LES CORPUS ANNOTES ?

La conjoncture actuelle tient, semble-t-il, à la rencontre[2] d'une tradition anglo-saxonne de linguistique descriptive s'appuyant sur les corpus électroniques et d'un profond changement de cap en traitement automatique du langage naturel[3] (désormais TALN[4]). Cette convergence apparente cache de profondes divergences sur la nature des données langagières à constituer et sur leur utilisation.

2.1 La linguistique descriptive anglo-saxonne et ses questions

Le rejet de principe, formulé par N. Chomsky dès 1957, du recours aux corpus au profit de l'appel à l'intuition du locuteur natif a relégué dans les limbes les travaux de linguistique quantitative et les études empiriques de données attestées. C'est, du moins, l'impression qui domine quand on se retourne sur les quarante dernières années de l'histoire de la linguistique.

Cette image est partiellement fausse. Dans le monde anglo-saxon, où l'empirisme bien compris garde toujours quelque attrait, parallèlement aux mutations des modèles chomskyens et de leurs avatars, s'est progressivement affirmée une linguistique faisant appel de plus en plus systématiquement à des corpus électroniques pour développer, à partir des « faits » rassemblés, des dictionnaires et des grammaires descriptives[5], mais aussi pour tester des hypothèses, confronter un modèle postulé aux réalisations effectives (Aarts, 1990). C'est le courant des linguistiques « de corpus » ou « sur corpus », en anglais *corpus linguistics*. Cette utilisation de corpus annotés, de grande taille, variés et assortis d'outils d'exploration puissants, permet d'observer plus finement les phénomènes et remet en question une partie des postulats de la linguistique.

Tout d'abord, la diversité même des corpus et le fait que certains d'entre eux ont été constitués pour rendre compte des registres et des genres langagiers permettent des études approfondies de la variation langagière. Il est possible d'étudier dans le détail, en dépassant les caractérisations trop globales, et donc caricaturales, l'opposition entre oral et écrit, l'organisation globale des textes, mais aussi les

[2] Notre analyse est proche de celle de M.-P. Péry-Woodley (1995).
[3] Cette dénomination est un calque maladroit de l'anglais *NLP (Natural Language Processing*). Elle pâtit de l'hésitation entre *langue* et *langage* pour la traduction de *language*. Rappelons qu'on entend par « langage naturel » une langue de communication, par opposition aux langages formels (notations logiques) et aux langages artificiels (langages de programmation). Comme le soulignait A. Guillet, la langue française marque la distinction entre les deux ordres langagiers. On dit *Il parle (le) verlan*, mais pas **Il parle (le) Prolog*.
[4] On se reportera à (Fuchs *et al.*, 1993) pour une présentation générale des domaines et des techniques du TALN.
[5] C'est le cas du *Survey of English Usage* de R. Quirk et de (Quirk *et al.*, 1985).

contrastes socio-linguistiques.

L'examen des corpus pose ensuite la question de l'articulation de la performance et de la compétence. Aux dires de G. Sampson (1994, p. 180) : « la linguistique de corpus prend le langage comme elle le trouve. » Le corpus ***Mitterrand1*** (Labbé, 1990, p. 95) présenté *infra* en 7.2.2 comprend par exemple l'énoncé suivant : « Moi, je suis de la France. Je ne dis pas : je suis la France. Je suis de la France. Toutes mes pensées, toutes mes façons d'être, toutes mes sensations, toutes mes vibrations, elles sont de la France[6]. » Plusieurs des constructions qu'emploie ici F. Mitterrand paraissent nettement a-grammaticales. Il ne s'agit pourtant pas d'un lapsus mais d'un choix délibéré, comme le prouvent les reprises. Si, comme l'affirme J.-C. Milner (1989, p. 55) : « [...] l'activité grammaticale ne consiste pas à enregistrer les données de langue ; elle consiste à émettre sur ces données un jugement différentiel », c'est-à-dire à isoler « l'impossible de langue » (*ibid.*), les linguistiques de corpus se trouvent confrontées à un éventail de réalisations langagières qui remet en cause les distinctions tranchées entre acceptable et non-acceptable.

Troisièmement, les corpus peuvent rassembler des énoncés sur lesquels l'analyste n'est pas forcément à même de porter des jugements d'acceptabilité. C'est le cas par exemple pour des corpus de langues mortes (Ancien Français, Anglais médiéval, etc.). Mais c'est aussi le cas pour des corpus de langues de spécialité, pour lesquels une partie des contraintes syntaxiques et sémantiques restent opaques à qui n'est pas « du domaine ». L'examen des régularités rencontrées au sein du corpus est alors un moyen, parfois le seul, de reconstituer la « grammaire » sous-jacente.

Enfin, même lorsqu'il s'agit d'un état de langue correspondant à la compétence langagière de l'analyste, un corpus permet d'apprécier l'importance relative des différentes réalisations. Certaines constructions, par exemple, sont extrêmement fréquentes, d'autres rares ou exceptionnelles. On peut penser que de tels décalages ne concernent pas vraiment la linguistique en tant que telle. Ce serait peut-être la position de J.-C. Milner (1989, p. 34) : « [...] toutes les questions que soulève la science du langage, dans toutes ses versions, sont des questions fines ; dès qu'elle dépasse la banalité, une proposition de linguistique concerne peu de données à la fois et elle y fait apparaître généralement ce que l'opinion courante tiendrait pour des détails. ». On peut aussi chercher à articuler les règles et le poids comparé des différentes régularités observées. Dans cette conception, les règles ne sont pas toutes sur le même plan : certaines sont centrales, d'autres périphériques. Les règles changent alors de statut. C'est une vision probabiliste de la grammaire (Sueur, 1982, p. 148-150).

2.2 Le changement de cap en TALN

La tradition des linguistiques de corpus a reçu ces dernières années un appui vigoureux et inattendu de la communauté du TALN, qui a donné un nouvel essor à la constitution et à l'utilisation de corpus annotés[7]. Cet appui découle de la prise de conscience progressive d'une inadéquation relative des paradigmes utilisés pour le TALN. En effet, la sophistication des formalismes utilisés ne débouche pas toujours

[6] Intervention radio-télévisée du 2 mars 1986.

[7] Ce renfort est souligné comme une heureuse surprise par un linguiste descriptiviste connu, G. Leech (1991, p. 20) : « Nous sommes maintenant dans une position où la recherche basée sur corpus a vraiment décollé, non seulement comme un paradigme d'investigation linguistique reconnu mais comme une contribution clé pour le développement de logiciels de traitement du langage naturel. La recherche [...] va probablement susciter non seulement l'attention des universitaires mais le financement industriel et public qui sera nécessaire si l'on veut obtenir les progrès souhaités. »

sur des systèmes de traitement fiables et efficaces. Deux causes sont généralement avancées. Tout d'abord, un système de TALN a besoin de ressources (dictionnaires, grammaires) à la fois très vastes (en nombre d'entrées lexicales et de règles) et très détaillées (concernant les conditions syntaxiques d'emploi des mots, par exemple). Les ressources actuelles sont notoirement insuffisantes, surtout en ce qui concerne la finesse de description. En second lieu, leur amélioration, semble-t-il, n'est ni uniquement ni même principalement à chercher dans des nouvelles études « en chambre » mais plutôt dans l'observation des larges ensembles de données textuelles qui sont maintenant disponibles. Il s'agit en fait d'un changement profond de paradigme. Jusque là, l'objectif des recherches en TALN et en Intelligence Artificielle était avant tout de « modéliser », de formaliser le savoir humain, de dégager les règles sous-jacentes. C'est pourquoi les méthodes utilisées en TALN étaient alors largement « symboliques », c'est-à-dire fondées précisément sur des règles[8]. M. Liberman (1991) résumait ainsi le courant anti-empirique, anti-numérique et pro-symbolique des vingt dernières années : « Compter était précisément considéré comme n'étant pas une tâche appropriée pour une personne de qualité. » L'observation de données langagières en très grande quantité et le traitement de flux d'informations aussi importants que ceux qui circulent aujourd'hui sur le réseau Internet conduisent inéluctablement à recourir à des approches quantitatives ou à marier approches symboliques et approches quantitatives.

C'est donc à un véritable changement de cap que nous assistons actuellement. Les enjeux industriels sont considérables. Ce n'est donc pas un hasard si les initiatives de constitution de corpus annotés et de ressources langagières de grande taille ont reçu dans le monde anglo-saxon des soutiens financiers importants, du secteur privé (édition), mais aussi de la puissance publique. La mise dans le domaine public de ces nouvelles ressources apparaît comme la condition *sine qua non* pour que les chercheurs et les industriels puissent progresser efficacement à partir de ces sources de connaissances.

Dans la communauté du TALN, l'accent est mis sur les très vastes ensembles de données textuelles (des centaines de millions de mots), l'objectif étant, comme le soulignent K. Church et R. Mercer (*ibid.*, p. 1) : « une couverture large (bien que peut-être superficielle) de texte tout-venant, plutôt qu'une analyse en profondeur de domaines (artificiellement) restreints. » Ce sont des traitements automatiques du langage ancrés fortement dans des données attestées (*data-intensive approach to language*) qui sont visés.

3. CHOIX TERMINOLOGIQUES

Nous employons le mot *corpus* dans une acception assez restreinte empruntée à J. Sinclair (1996, p. 4) : « Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour

[8] Deux signes, parmi bien d'autres, de cette prééminence. A la fin des années quatre-vingts, paraissaient deux « sommes » sur le TALN. La première (Gazdar et Mellish, 1989) présentait les formalismes d'unification et écartait dans l'introduction toute quantification : « Comme tous ceux qui comptent les moutons le savent bien, compter est une tâche parfaitement ennuyeuse. Même les premiers ordinateurs comptaient vite et bien sans en mourir d'ennui ». G. Gazdar et C. Mellish ajoutaient à propos des index et concordances : « Aujourd'hui de tels travaux continuent sous la rubrique 'linguistique, littérature et ordinateur' mais ne relèvent plus de la linguistique computationnelle. » B. Partee et ses collègues, dans leur vaste présentation des modèles mathématiques pour la linguistique (1990), ne mentionnaient qu'une fois en 613 pages les modèles statistiques et probabilistes ... pour dire qu'ils ne seraient pas abordés.

servir d'échantillon du langage. » Nous précisons cette optique au chapitre VI. À cette aune, nombre de ressources textuelles perdent cette dénomination. Il s'agit souvent de collections ou de rassemblements de textes électroniques plutôt que de corpus à proprement parler.

Nous empruntons au québecois le terme *parsage* (*parsing*) pour désigner l'analyse syntaxique automatique et le mot *parseur* (*parser*) pour le programme qui effectue cette opération.

En recherche d'information, la *précision* représente la proportion de réponses pertinentes données par rapport au total des réponses extraites. Le *rappel* est la proportion des réponses pertinentes extraites par rapport au total des réponses pertinentes possibles. Le *silence* correspond alors les réponses pertinentes non extraites. Le *bruit* renvoie aux informations non pertinentes produites.

Par difficulté à trouver une expression satisfaisante, nous parlons parfois d'*annotation manuelle*, par opposition à une *annotation automatique*, c'est-à-dire effectuée par un programme. L'annotation n'est jamais vraiment « manuelle » : des programmes spécifiques ont pour objectif de faciliter le travail de la personne qui annote (l'*annotateur* ou l'analyste) voire de vérifier partiellement la cohérence des informations qu'elle fournit. Inversement, l'annotation automatique est souvent précédée ou suivie d'interventions humaines[9].

Annoter revient à regrouper sous un même chef, un même *type*, des réalisations linguistiques distinctes, ses *occurrences*. C'est le lemme pour les flexions d'un mot : *grand* pour *grand*, *grands*, *grande*, *grandes*. Il peut s'agir d'une classe plus abstraite. Les suites de mots *le président de la république* et *le livre des Rois* sont deux occurrences du type syntagme nominal, tout comme *je*, *ici* et *maintenant* constituent trois occurrences du type embrayeur.

Signalons enfin que nous employons souvent le mot *ambiguïté* pour des situations où un locuteur n'en perçoit pas. Le fait de dire que *pomme de terre* peut éventuellement être ambigu dans *Il sort les pommes de terre* paraît relativement raisonnable. Il n'en va pas de même pour *Il prend les pommes de terre*. Pourtant, les programmes de traitement ne disposent pas toujours des connaissances qui leur permettraient de choisir dans de tels cas. Il est d'usage en TALN de parler d'ambiguïté pour ces situations. C'est cet usage que nous suivons. La *désambiguïsation* consiste à choisir entre un certain nombre de possibilités.

4. NOTATIONS

Les corpus et les ressources textuelles sont cités par leur nom seul[10], sans déterminant, en gras italique. Nous parlons de ***Brown*** et non du corpus Brown ou du Brown, à la fois pour limiter le retour du mot *corpus*, déjà bien suffisamment à l'honneur dans ces pages et pour éviter de statuer sur l'adéquation de la notion, telle que nous l'entendons, à l'ensemble textuel considéré.

Les mentions des corpus, des ressources textuelles, des auteurs et des notions sont rassemblées dans un même index.

Les termes techniques (avec éventuellement leur correspondant anglais entre parenthèses[11]) sont détachés en italiques lors de leur première utilisation. Ils sont

[9] Cf. chapitre VIII.
[10] Il s'agit souvent d'un acronyme (***Susanne***, ***Archer***) ou du lieu ou de l'institution à l'origine du corpus (***Brown***), ou d'un mélange des deux (***LOB*** : London-Oslo-Bergen).
[11] Sauf dans quelques cas bien spécifiques, comme parsage, nous cherchons à éviter les anglicismes.

repris dans l'index.

Les crochets servent à isoler des suites de traits linguistiques, qui sont mis entre accolades : [{nom commun}{adjectif relationnel}] désigne l'enchaînement d'un nom commun et d'un adjectif relationnel.

Les exemples extraits de corpus et les sorties d'analyseurs sont signalés par un changement de police comme dans {adjectif relationnel}.

5. ORIENTATION DE L'OUVRAGE

Devant la multiplicité des points de vue possibles sur cette conjoncture nouvelle et les travaux qui en sont issus, nous précisons les parti-pris qui sont les nôtres dans les pages qui suivent.

5.1 L'écrit au travers de corpus enrichis de langues vivantes

Nous avons mis l'accent sur les corpus relevant de l'écrit[12]. Les corpus d'oral transcrit sont encore rares : la transcription proprement dite, les choix qu'elle entraîne, les coûts qu'elle suppose freinent leur développement, même si celui-ci semble s'accélérer dans les dernières années[13]. On dispose d'un recul moindre pour ce domaine que pour celui de l'écrit. Il nous semble aussi que l'oral impose des niveaux de description et des outils théoriques partiellement éloignés de ceux qui sont traditionnellement utilisés pour l'écrit[14]. Ce cadre théorique nous fait défaut. Il nous a semblé préférable de laisser d'autres en parler mieux que nous.

De nombreux textes latins et grecs sont disponibles sous forme électronique. Nous ne parlerons cependant pas de ces corpus de langues mortes. Nous centrons en effet notre analyse sur les langues vivantes ainsi que sur les états anciens de ces langues (l'ancien et le moyen français, par exemple).

Par *corpus enrichis* ou *annotés*, nous entendons des corpus dans lesquels les séquences de caractères qui constituent les mots sont assorties d'autres informations[15] : lemmes, étiquettes morpho-syntaxiques, sémantiques, arbres syntaxiques, apparat critique, etc. Nous ne retenons pas les corpus « nus », c'est-à-dire faits de mots seuls, sans annotation, sauf à l'occasion pour montrer l'écart entre les analyses selon le niveau d'information disponible ou dans le cas de corpus d'états anciens des langues actuelles.

Nous présentons cependant d'autres ressources textuelles qui ne sont pas des corpus annotés mais qui représentent tout de même une source d'information précieuse. C'est le cas des versions électroniques de dictionnaires papier ou de

[12] Certains des corpus mentionnés, comme ***BNC***, comprennent en fait une partie d'oral. Deux des corpus français utilisés sont partiellement ou totalement de l'oral transcrit : ***Mitterrand1***, mais aussi les lettres et les compte-rendus dans ***Menelas***, qui sont dictés. Certaines recherches présentées (celles de Biber, par exemple, résumées au chapitre I) font appel à l'oral. Mais c'est une dimension que nous laissons à chaque fois dans l'ombre.

[13] On distingue en outre corpus d'oral et corpus de parole (Sinclair, 1996, p. 8-9). Les premiers servent aux linguistes et reposent sur des transcriptions associant éventuellement alphabet phonétique et signes spécifiques pour noter la prosodie, etc. Les seconds relèvent de la communauté de la reconnaissance de la parole et restent plus proches d'enregistrements.

[14] Cf. (Blanche-Benveniste, 1997).

[15] G. Sampson, « père » de ***Susanne*** (cf. infra), distingue conventions d'annotation *(annotation scheme*) et système d'annotation (*annotation system*) : méthode qui met en œuvre ces conventions. La méthode peut être manuelle ou automatique.

thesaurus. C'est le cas aussi des textes alignés, où l'un des textes est la traduction de l'autre. Aujourd'hui, on ne dispose plus seulement de corpus annotés préalablement, mais d'outils permettant de traiter de nouveaux textes et de constituer de nouveaux corpus enrichis. Ces outils d'annotation (étiqueteurs, analyseurs syntaxiques ...) retiennent aussi notre attention.

5.2 Les corpus, les ressources et les recherches de langue anglaise

Qu'on ne voie ni une anglophilie excessive ni un engouement coupable pour la modernité américaine dans l'attention que nous accordons aux corpus, aux ressources en anglais ou en américain et aux travaux qui s'en servent, anglo-saxons eux aussi pour la plupart.

Nécessité fait loi. Les corpus enrichis sont aujourd'hui majoritairement de langue anglaise ou américaine[16] ... même lorsqu'ils sont développés dans des pays extérieurs au monde anglo-saxon : c'est le cas du corpus de Nimègue aux Pays-Bas ainsi que d'***Helsinki***. Les travaux qui utilisent ces ressources paraissent avant tout dans des colloques, des revues et des livres anglais ou américains. Les outils d'annotation et les dictionnaires électroniques sont aussi majoritairement développés pour la langue anglaise ou américaine. Cet état de fait résulte à la fois de l'ancienneté d'une tradition anglo-saxonne de linguistique descriptive appuyée sur des corpus et de la place prééminente de l'anglais et de l'américain dans les projets de TALN depuis les débuts de ces recherches.

La francophonie s'engage dans ce mouvement, avec un certain retard et une réticence certaine à mettre dans le domaine public des ressources comme des corpus étiquetés et des étiqueteurs. À terme, ces ressources n'en seront pas moins disponibles. Nous avons donc complété un exposé essentiellement consacré à des travaux anglo-saxons par la présentation de corpus annotés de langue française et d'outils destinés à notre langue.

5.3 Un point de vue aux frontières de la linguistique

Nos domaines de spécialité (analyse syntaxique automatique, sémantique formelle et statistique textuelle) nous situent aux frontières de la linguistique. C'est peut-être un regard oblique que nous portons sur les recherches dont nous rendons compte. Nous ne prétendons pas juger la pertinence linguistique des études que nous avons retenues. Nous cherchons à mettre en évidence les grandes tendances que nous percevons. Il ne nous semble d'ailleurs pas possible de pouvoir prétendre faire état d'un ensemble représentatif des travaux relevant des linguistiques de corpus. Il faudrait une culture linguistique à la fois extrêmement vaste et très approfondie sur certains points pour appréhender et évaluer la multiplicité des travaux linguistiques à partir de corpus. Nous espérons tout de même que notre insertion dans des projets interdisciplinaires nous aura permis de percevoir (et de faire sentir) l'aspect séminal de certaines recherches. Peut-être notre regard oblique se révèlera-t-il rafraîchissant.

[16] Nous distinguons l'anglais et l'américain dans ce livre, dans la mesure précisément où l'existence de corpus comparables comme ***LOB*** et ***Brown*** a permis des études contrastives sur ce point, comme (Mair, 1995).

5.4 La diversité des publics concernés

S'il met l'accent sur les recherches linguistiques s'appuyant sur des corpus annotés, cet ouvrage n'est pas uniquement destiné aux linguistes. La didactique des langues est aussi concernée. Les corpus représentent des ressources importantes pour l'apprentissage des langues : phénomènes collocatifs et phraséologie, micro-syntaxe des entrées lexicales, étude des langues de spécialité, typologie des textes. Nous abordons tous ces aspects. La lexicographie, en particulier spécialisée (la terminologie), commence déjà à utiliser les méthodes et les outils qui sont présentés ici. L'analyse de contenu peut tirer profit des nouveaux outils de traitement. Nous montrons ainsi sur un corpus de réponses à des questions ouvertes l'emploi d'étiquettes morpho-syntaxiques pour contraster plus finement les styles sociaux des locuteurs. Les chercheurs du TALN, qui peuvent relativement facilement se procurer « du » texte électronique, trouveront dans ces pages des indications méthodologiques sur la constitution de corpus, en particulier sur l'influence des genres textuels.

Nous parlons de linguistiques de corpus au pluriel pour souligner cette diversité d'approches[17].

6. DEMARCHE SUIVIE

L'ouvrage se divise en trois parties. Nous partons des corpus annotés et des autres ressources textuelles disponibles. Nous abordons ensuite d'autres dimensions du travail sur corpus : l'étude du sens, celle de la diachronie, les textes alignés. Nous finissons par les problèmes méthodologiques et techniques, plus abstraits pour les premiers, plus éphémères pour les seconds.

Les renvois bibliographiques, nombreux, témoignent de l'intense activité de recherche et de développement autour des corpus électroniques. Ils comprennent des actes de conférence et même des rapports techniques : la recherche est active dans ce domaine.

6.1 Les corpus annotés et leurs utilisations

Le chapitre I aborde les corpus étiquetés : des étiquettes morpho-syntaxiques sont associées aux mots. Le chapitre II traite des corpus arborés : des représentations syntaxiques décorent les phrases.

Au sein de chacun de ces chapitres, nous présentons d'abord rapidement le niveau d'annotation concerné. Les corpus présentés à la fin de cette introduction sont sollicités pour des exemples où nous respectons les lourdeurs des notations existantes. Nous essayons en même temps de fournir une représentation unifiée pour chaque niveau d'annotation de manière à pouvoir comparer les formats effectifs utilisés, ces derniers étant extrêmement variés. Les différences de notations empêchent en effet souvent de percevoir les divergences et les convergences

[17]Nous rejoignons M.-P. Péry-Woodley (1995) : « Le fait que n'existe pas en français un terme unificateur [comme *corpus linguistics*] a pour conséquence que rien ne vient cacher la diversité des objectifs et des méthodes des différents utilisateurs de corpus. »

réelles. Dans un deuxième temps, nous développons quelques exemples de recherches linguistiques rendues possibles par ce niveau d'annotation et qui paraissent particulièrement prometteuses. Par ces exemples, nous voulons montrer d'emblée ce que peuvent apporter les différents niveaux d'annotation possibles d'un corpus, sans que les problèmes techniques viennent troubler la perception des enjeux.

Le chapitre III décrit d'autres ressources textuelles importantes : les ressources lexicales sous forme électronique.

6.2 Dimensions transversales

Le chapitre IV, consacré au volet sémantique, montre comment extraire des connaissances lexicographiques de corpus ou désambiguïser le sens des mots en contexte.

Le chapitre V présente l'utilisation de corpus dans une perspective diachronique, sur la longue durée ou au contraire sur des périodes courtes. Il indique les difficultés propres de la constitution de corpus historiques et les précautions méthodologiques nécessaires lors de leur utilisation.

Le chapitre VI décrit les textes alignés : un texte écrit dans une langue est mis en parallèle avec sa version dans une ou plusieurs autres langues.

6.3 Méthodologies et techniques

La dernière partie regroupe les réflexions méthodologiques et les informations techniques.

La compréhension préalable des études utilisant des corpus rend plus tangibles les enjeux de la constitution d'un corpus et les choix méthodologiques qu'elle nécessite, en particulier en ce qui concerne les normes destinées à faciliter l'échange et la réutilisation des données textuelles (SGML, TEI). C'est l'objet du chapitre VII.

En essayant d'éviter l'hermétisme, bien conscients que c'est probablement le point sur lequel les évolutions sont les plus rapides et les plus difficiles à anticiper, nous présentons au chapitre VIII les techniques d'étiquetage et d'analyse syntaxique, celles d'annotation sémantique, ainsi que le « toilettage » et la segmentation des données textuelles.

Le chapitre IX présente rapidement la quantification des faits langagiers.

7. PRINCIPAUX CORPUS CITES

Les corpus annotés sont aujourd'hui légion, et nous ne saurions prétendre en dresser la liste. Cependant, certains d'entre eux sont devenus canoniques, soit du fait des méthodes employées pour les constituer et les annoter (c'est le cas de ***Susanne***, par exemple), soit en raison des études linguistiques qui les ont utilisés. L'index renvoie aux passages où ces deux aspects sont évoqués. Ce sont ces corpus anglais et américains que nous présentons. Ils sont en général disponibles

pour la recherche universitaire.

7.1 Corpus anglais ou américains

- ***Brown*** Ce corpus étiqueté d'un million de mots a été mis au point en 1979 par W. Francis et H. Kucera, à l'université Brown (USA). Il comprend 500 extraits de 2 000 occurrences chacun provenant de textes américains publiés en 1961 et relevant de 15 « genres » : reportage, écrits scientifiques et techniques, etc. Il a été soigneusement étiqueté. Par sa mise dans le domaine public, il a joué un rôle moteur dans le renouveau des études sur corpus.
- ***LOB*** (Lancaster-Oslo-Bergen) Ce corpus étiqueté a été conçu comme l'équivalent anglais de ***Brown***. Il comprend également 1 million de mots sélectionnés selon les mêmes critères mais à partir de textes anglais publiés en 1961.
- ***Susanne*** Ce corpus de 128 000 occurrences annoté sous la direction de G. Sampson (1994, 1995) est constitué de 64 extraits de 2 000 occurrences chacun pris dans ***Brown***. Il comprend des reportages, des textes littéraires (romans, biographies, mémoires), des écrits scientifiques et techniques et enfin des textes de fiction. La particularité de ***Susanne*** est que chaque phrase est assortie d'un arbre syntaxique très détaillé, associant des étiquettes catégorielles et des étiquettes fonctionnelles.
- ***London-Lund*** Ce corpus étiqueté (Svartvik *et al.*, 1982) totalise 435 000 mots d'anglais parlé, répartis en 87 extraits de 5 000 occurrences de locuteurs adultes ayant fait des études. Il inclut conversations, y compris téléphoniques, conférences et cours, commentaires radiophoniques, etc. Il comprend de nombreuses informations prosodiques (pauses, limites, etc.).
- ***Lancaster/IBM Treebank*** Ce corpus arboré (Black *et al.*, 1993) rassemble 1 million de mots de l'agence Associated Press, 1 million de mots issus des débats du parlement canadien, 250 000 mots de APHB (American Printing House for the Blind), 800 000 mots de manuels IBM. Il est muni d'une annotation syntaxique limitée : parenthésage et étiquetage des constituants.
- ***Helsinki*** C’est un corpus pour l’étude diachronique de l’anglais. Il comprend 1,5 millions de mots non annotés, couvrant la période allant de l’année 750 à 1700, répartis en 11 périodes et différents types de textes (Kyto, 1993a ; 1993b).
- ***Archer*** C’est un corpus pour l’étude diachronique de l’anglais et de l’américain. Il comprend 1,7 million de mots non annotés, de l’année 1650 à 1990, répartis en périodes de cinquante ans et en genres (journaux intimes, fiction, écrits journalistiques, médecine, science, décisions de justice, théâtre, sermons, etc.).
- ***BNC*** (British National Corpus) Ce corpus étiqueté de 100 millions de mots mêle oral (10 %) et écrit (textes de fiction à partir de 1960 et textes « informatifs » à partir de 1975). Les échantillons sont représentatifs d'une grande diversité de situations langagières, mais sans organisation par thèmes, registres ou genres (Burnard, 1995).
- ***Penn Treebank*** Ce corpus arboré (Marcus *et al.*, 1993) comprend 4 millions de mots issus de sources diverses : Manuels IBM, ***Brown***, Department of Energy, Department of Agriculture, textes littéraires, Library of America, oral transcrit, DARPA Air Travel Information System, informations financières, Dow Jones.

7.2 Corpus français

Nous ajoutons trois corpus français annotés. Ils associent langue spécialisée (***Menelas***) et langue générale (***Mitterrand1***, ***Enfants***). Ils ne sont pas dans le domaine public.

- ***Menelas*** Ce corpus étiqueté et partiellement arboré, de 84 839 occurrences et 6 191 formes différentes, a été rassemblé pour le projet européen Menelas (Zweigenbaum, 1994) de compréhension de comptes rendus d'hospitalisation. Il concerne les maladies coronariennes[18]. Il réunit un extrait de manuel médical, des comptes rendus d'hospitalisation et des lettres des médecins hospitaliers à leurs collègues non hospitaliers à propos de patients communs.
- ***Mitterrand1*** Ce corpus étiqueté et lemmatisé[19] regroupe les interventions radio-télévisées de F. Mitterrand au cours de son premier septennat. Il a été constitué par D. Labbé (Institut d'Etudes Politiques de Grenoble). Il compte 305 124 occurrences et 9 309 formes. La qualité du travail d'étiquetage et de lemmatisation, ainsi que la minutie de la vérification font de ce corpus de taille moyenne un excellent observatoire de la langue « générale » (par opposition par exemple à ***Menelas*** décrit *supra*).
- ***Enfants*** Ce corpus[20] est constitué de réponses à la question : « Quelles sont les raisons qui, selon vous, peuvent faire hésiter une femme ou un couple à avoir un enfant ? » Cette question a été posée en 1981 à 2 000 personnes représentant la population des résidents métropolitains de 18 ans et plus lors d'une enquête effectuée par le Centre de Recherches et de Documentations sur la Consommation (CREDOC), sous la direction de L. Lebart, sur les conditions de vie et les aspirations des Français. Ce corpus comprend 15 523 occurrences (ponctuation non comprise) et 1 305 formes. Chaque réponse est précédée d'indications sociologiques sur la personne interrogée (sexe; âge,niveau de diplôme, etc.).

[18] Il a aussi servi de banc d'essai à un certain nombre de méthodes d'acquisition automatique ou assistée de terminologies scientifiques et techniques.
[19] Il a été étudié dans une perspective politologique (Labbé, 1990).
[20] Il a été étudié au niveau des « mots » dans (Lebart et Salem, 1994). Une fois lemmatisé, étiqueté et corrigé, il a été analysé dans (Habert et Salem, 1995).

PREMIERE PARTIE

LES CORPUS ANNOTES ET LEURS UTILISATIONS

CHAPITRE I

LES CORPUS ETIQUETES

Etiqueter un texte, c'est une forme d'annotation dans laquelle on associe à des segments de texte, le plus souvent les « mots », une ou plusieurs étiquettes, le plus leur catégorie grammaticale voire leur lemme.

Dans la première section, nous donnons de brefs exemples de corpus étiquetés et nous définissons les types d'étiquetage rencontrés. Un premier exemple d'utilisation de corpus étiquetés (section 2) repose sur un étiquetage approfondi d'une partie seulement du corpus. Il vise à mettre en évidence de manière inductive une typologie des textes sur la base des corrélations observées entre les traits linguistiques retenus. Un second exemple (section 3) fait appel à un étiquetage complet mais fruste (la partie du discours et quelques renseignements morphologiques). Cet étiquetage permet de contraster les « parlures » qui coexistent dans le corpus étudié. Nous abordons en section 4 l'utilisation d'étiqueteurs ou de corpus étiquetés et en section 5 les enjeux théoriques des recherches rendues possibles par ce niveau d'annotation.

8. DEFINITIONS

Commençons par trois brefs exemples, qui donnent un aperçu de la diversité des étiquetages effectifs ... comme de leur manque de lisibilité et de clarté.

8.1 Exemples

8.1.1.1 Enfants

Les réponses fournies par les personnes interrogées :

> Les difficultés financières et matérielles.
>
> Je ne sais pas, les gens sont égoïstes, peut-être.

sont lemmatisées et étiquetées (cf. 3.2) de la manière suivante :

> <S01=23> le les {DETDEF} difficulté difficultés {NOMFP} financier financières {ADJFP} et et {CCOORD} matériel matérielles {ADJFP} . . {PONCT-FORTE}
>
> <S01=31> je je {PROPERS} ne ne {ADVNEG} savoir sais {VIPR1S} pas pas {ADVNEG} , , {PONCT-FAIBLE} le les {DETDEF} gens gens {NOMMP} être sont {VIPR3P} égoïste égoïstes {ADJMP} peut-être peut-être {ADV} . . {PONCT-FORTE}

Chaque réponse commence par des renseignements sur l'interviewé : son âge (en deuxième position après S01= : 1 renvoie à inférieur à 30 ans, 2 à entre 30 et 50 ans, 3 à au delà de 60 ans) et son niveau d'étude (en première position après S01= : 1 = sans, 2 = baccalauréat, 3 = études supérieures). Puis chaque mot, précédé de son lemme, est suivi de sa catégorie morphosyntaxique entre accolades (NOMMS = nom masculin singulier, par exemple).

8.1.1.2 Mitterrand1

Le fragment suivant est extrait de l'émission de TF1 *Ça nous intéresse, Monsieur le président* du 2 mars 1986 :

> [...] moi, je suis de la France - je ne dis pas : je suis la France - [...]

Il est codé de la manière suivante par D. Labbé (1990) :

moi,je,5
","","","p"
je,je,5
suis,être,11
de,de,81
la,le,7
France,France,22
-,-,p
je,je,5
ne,ne,6
dis,dire,11
pas,pas,6
:,:,p
je,je,5
suis,être,11
la,le,7
France,France,22

Le texte annoté est constitué d'une série de triplets comme suis,être,11 : le mot, le lemme, la catégorie, représentée par un nombre. Les trois informations sont séparées par des virgules.

8.1.1.3 Susanne

La phrase :

DAN MORGAN TOLD HIMSELF HE WOULD FORGET Ann Turner[21] :

est représentée ainsi :

N01:0010b	NP1m	DAN	Dan	[O[S[Nns:s.
N01:0010c	NP1s	MORGAN	Morgan	.Nns:s]
N01:0010d	VVDv	TOLD	tell	[Vd.Vd]
N01:0010e	PPX1m	HIMSELF	himself	[Nos:i.Nos:i]
N01:0010f	PPHS1m	HE	he	[Fn:o[Nas:s.Nas:s]
N01:0010g	VMd	WOULD	will	[Vdc.
N01:0010h	VV0v	FORGET	forget	.Vdc]
N01:0010i	NP1f	Ann	Ann	[Nns:o.
N01:0010j	NP1s	Turner	Turner	.Nns:o]Fn:o]S]
N01:0010k	YF	+.	-	.

Le texte est ici présenté sous la forme d'un tableau : à un mot du texte de départ correspond une ligne. Chaque ligne fournit une suite de champs. Ici pour la troisième ligne :

N01:0010d	VVDv	TOLD	tell	[Vd.Vd]

- une référence : le nom du fichier dont provient cet extrait (N01) et un numéro de ligne au sein de ce fichier : 0010d ;
- une indication d'édition : le tiret indique que le texte n'a pas été corrigé à cet endroit ;
- une catégorie : VVDv ;
- la forme fléchie telle qu'on la rencontre dans le corpus : told ;
- le lemme correspondant : tell ;
- la structure syntaxique dans laquelle s'insère le mot[22] : [Vd.Vd] indique que ce mot est la tête d'un groupe verbal. Le point signale l'endroit où le mot et sa catégorie doivent s'insérer. C'est l'équivalent de [Vd [VVdv told]].

8.2 L'inévitable éparpillement des étiquetages

Les exemples donnés manifestent la diversité en taille et en visée des jeux d'étiquettes et des stratégies d'étiquetage sous-jacentes. Cette

[21] Les majuscules sont dans le texte de départ.
[22] Nous reviendrons sur ce dernier champ au chapitre suivant, consacré aux corpus arborés : cette annotation syntaxique n'est généralement pas considérée comme faisant partie de l'étiquetage à proprement parler.

diversité tient à l'utilisation envisagée du corpus mais aussi à son mode d'étiquetage (manuel ou automatique) ainsi qu'à l'absence de consensus sur certains catégories ou sur leur extension.

L'expérience montre qu'un groupe d'annotateurs n'est pas forcément cohérent dans les étiquettes qu'il attribue manuellement à un texte. Il en va de même pour un même individu au fil du temps.

J. Véronis et L. Khouri soulignent (1995, p. 235) le fait que les jeux d'étiquettes ne sont généralement pas comparables, ce qui retarde l'évaluation ou la combinaison des étiqueteurs et des étiquetages. Pour reprendre Leech et ses collègues (1994, p. 51) : « il n'y a pas de 'meilleur jeu d'étiquettes', [...] dans la pratique la plupart des jeux d'étiquettes constituent plutôt des compromis entre la finesse de la description linguistique et ce qui peut être attendu, pour des raisons pratiques, d'un système automatique d'étiquetage[23]. » On peut recourir à un jeu d'étiquettes important pour pouvoir distinguer aisément certains cas d'ambiguïté, quitte à se ramener à un jeu plus restreint une fois l'étiquetage opéré[24]. Inversement, sur certains points, le jeu d'étiquettes peut en rester à des distinctions relativement grossières, parce qu'il s'avère difficile d'obtenir, sur des subdivisions plus fines, un consensus de la part des personnes définissant l'ensemble d'étiquettes à utiliser (Greenbaum, 1993) ou parce que des catégories trop fines rendraient plus long et plus hasardeux le travail de correction manuelle des résultats de l'étiquetage automatique. Greenbaum (*ibid.*, p. 18) donne l'exemple de la distinction comptable / non comptable, importante pour les noms en grammaire anglaise, mais difficile à établir avec sûreté, *a fortiori* à automatiser. Il propose alors de s'en tenir à l'opposition, aisément détectable, entre singulier et pluriel. À charge pour ceux qui entendent précisément étudier la dimension comptable / non comptable d'annoter en conséquence leur corpus !

Par ailleurs, les jeux d'étiquettes correspondent aussi sur certains points à des divergences théoriques réelles. Il en va de même de la projection des catégories, soulignent J. Véronis et L. Khouri (*ibid.,* p. 237) : « Même si l'on est d'accord sur le jeu d'étiquettes, leurs extensions (c'est-à-dire l'ensemble des formes lexicales qu'elles couvrent) peuvent être différentes. Le problème est particulièrement aigu pour les catégories fermées, déterminants, pronoms, adjectifs indéfinis, etc., où l'on rencontre de très grosses différences d'appréciation dans les catégories, et ce dans la plupart des langues. » Comme l'indiquent Greenbaum et Yibin (1994, p. 35) : « l'identité des étiquettes [entre deux jeux] peut être trompeuse, dans la mesure où l'assignation des étiquettes peut être différente. » Ils citent le cas de l'étiquette adverbe qui est conservée par ***ICE*** (International Corpus of English) pour les adverbes utilisés comme modifieurs de noms (*then* dans *the then president*) mais que l'étiqueteur CLAWS remplacerait par l'étiquette adjectif. Dans les cas

[23] J.-P. Chanod et P. Tapanainen (1995a) indiquent ainsi qu'ils ont ignoré la distinction masculin / féminin en français pour les noms et les adjectifs, dans la mesure où cette distinction suppose l'utilisation de contextes larges *(une envie de soleil diffuse*) et où finalement, pour leurs objectifs (repérage de l'accord sujet / verbe et ambiguïté nom / verbe), elle joue un rôle mineur.

[24] C'est la pratique d'E. Tzoukermann et de ses collègues (1995) avec des jeux de 253 et 67 étiquettes respectivement.

de conversion, c'est-à-dire de passage d'une catégorie à une autre sans changements dérivationnels, doit-on attribuer la catégorie de départ ou celle d'arrivée ? Comment catégoriser par exemple *parler* dans la séquence « le parler vrai » : comme un infinitif ou comme un nom ?

8.3 Une représentation canonique

Les corpus étiquetés peuvent donc se présenter sous des formats variables : verticalement (comme ***Mitterrand1*** ou ***Susanne***) ou horizontalement (***Enfants***). Dans ces trois exemples, la nature des informations doit être déduite de l'usage de divers caractères qui prennent un sens particulier : crochets, point, virgule, accolades, passages à la ligne, ainsi que de la place où les informations figurent. La catégorie constitue la troisième colonne de ***Susanne*** et de ***Mitterrand1*** et elle occupe la troisième position de chaque triplet pour ***Enfants***.

On peut figurer ainsi le décodage de l'étiquetage d'un mot annoté dans ***Mitterrand1*** :

mot	séparateur de champ	lemme	séparateur de champ	catégorie	séparateur de triplet
suis	,	être	,	11	passage à la ligne

Pour faciliter la récupération d'un champ donné et la transmission des corpus, on doit passer de ces indications positionnelles à une représentation logique, ce qui revient à isoler chaque type d'information et à lui donner un nom, soit avant cette information :

catégorie=verbe, lemme=être, forme=être

soit « autour » de cette information :

<catégorie>verbe</catégorie><lemme>être</lemme><forme>suis</forme>.

Cette dernière représentation, destinée à faciliter les échanges et réutilisations de corpus, repose sur des normes de balisage présentées au chapitre VI.

Ces conventions rendent explicite une représentation canonique de l'étiquetage. Les informations associées à un segment de texte peuvent en effet être représentées par une structure d'associations trait-valeur du type de celles utilisées par les formalismes syntaxiques contemporains[25]. Nous notons ces structures entre accolades, chaque trait étant séparé par le signe = de sa valeur à cet endroit et par une virgule du trait suivant. La ligne de ***Susanne*** donnée *supra*, abstraction faite du champ notant l'analyse syntaxique, se note alors :

[25] On se reportera à (Abeillé, 1993, p. 29-31) pour une présentation générale de ces structures et à (Ligozat, 1994, ch. 3 et ch. 5) pour un approfondissement formel.

{référence=N01:0010d, catégorie=VVDv, forme=told, lemme=tell}

et celle de Mitterrand1 :

suis,être,11

se transcrit ainsi :

{forme=suis, lemme=être, catégorie=11}

Comme les noms des traits sont fournis, on peut disposer les associations trait-valeur dans n'importe quel ordre. La version suivante de la ligne de de ***Mitterrand1*** est strictement équivalente à la précédente :

{catégorie=11,forme=suis,lemme=être}

Enfants ne fournit que la catégorie et le lemme, à côté des indications sur le diplôme et l'âge du locuteur. Ces indications pourraient être elles-mêmes ajoutées sous forme de traits attachés à chaque mot. Elles seraient alors « distribuées » au lieu d'être mises en facteur, ce qui donnerait, en format vertical :

{diplôme=baccalauréat, âge=60+, catégorie=DETDEF, forme=les, lemme=le}

{diplôme=baccalauréat, âge=60+, catégorie=NOMFP, forme=difficultés, lemme=difficulté}

De telles structures de traits sont « ouvertes » : il est toujours possible de rajouter des « dimensions » (par exemple des étiquettes sémantiques). On peut également enlever une partie des associations trait-valeur attachées à un « mot » et simplifier par là-même son étiquetage. On en verra un exemple dans la section 3.

Susanne fournit un trait référence identifiant de manière unique le mot examiné. Dans ***Mitterrand1***, il faut connaître le fichier dont provient l'occurrence. Le soin apporté par ***Susanne*** sur ce point peut paraître superflu. C'est pourtant en définitive sur cette identification univoque que repose la possibilité de vérifier les annotations portées sur un corpus ou les analyses qui en sont faites. Un autre chercheur peut se reporter exactement au bon endroit dans le texte de départ, examiner un contexte plus large, etc. C'est donc la condition *sine qua non* d'un travail collectif.

Si l'on adopte cette représentation canonique, on constate que le trait catégorie est utilisé différemment selon les cas. Par exemple, pour le mot *je*, la valeur de ce trait est 5, c'est-à-dire Pronom pour ***Mitterrand1*** et PROPERS pour ***Enfants***. Dans ce corpus, l'étiquette précise donc, de manière relativement transparente, le type de pronom dont il s'agit. On peut alors expliciter les composants d'une telle étiquette : {catégorie=pronom, type=personnel}. Il est fréquent que les étiquettes d'un corpus ne soient pas atomiques mais complexes : on doit les décomposer. C'est le cas pour ***Susanne***, où VVDv est en fait une abréviation pour : {categorie=verbe, temps=passé}. Développer ainsi les étiquettes complexes[26] facilite l'élagage

[26] Le projet européen MULTEXT de création de ressources linguistiques informatisées, monolingues et multilingues, et d'outils génériques d'annotation et d'exploitation de

ou l'enrichissement des traits attachés à un « mot ».

8.4 Types d'étiquetage

L'étiquetage peut être produit par un programme qu'on appelle un étiqueteur (*tagger*), ou bien résulter d'une annotation manuelle, ou bien provenir d'une combinaison des deux. Le traitement de gros volumes de textes rend cependant inéluctable le recours à un étiqueteur.

8.4.1 Etiquetage intégral ou partiel

Dans les exemples que nous avons fournis, chaque mot fait l'objet d'un étiquetage. On rencontre par ailleurs des textes étiquetés partiellement : les renseignements attachés à certains mots sont inexistants on incomplets. Il peut s'agir de limites purement techniques : l'étiqueteur utilisé bute sur des mots « inconnus », c'est-à-dire absents des dictionnaires qu'il utilise ou que ne résolvent pas les règles morphologiques qu'il emploie. Ou bien, face à un mot inconnu, l'étiqueteur fait des propositions moins précises que celles déclenchées par les mots répertoriés dans les dictionnaires employés.

L'étiquetage partiel peut aussi être visé en tant que tel. Un sous-ensemble des mots du texte est jugé pertinent pour la recherche envisagée, il est donc étiqueté, le reste est ignoré. Par exemple, si l'on entend étudier la répartition des marques de l'énonciation dans un corpus, on peut envisager un étiquetage limité aux mots retenus comme révélateurs sur ce point : embrayeurs, certains adverbiaux, indications temporelles et aspectuelles des verbes ...

8.4.2 Une étiquette ou plusieurs étiquettes

Un corpus étiqueté n'est pas forcément totalement « désambiguïsé », c'est-à-dire qu'un mot peut recevoir plusieurs étiquettes. Dans ***BNC***, à l'issue de l'étiquetage, demeurent un peu plus de 3 % de problèmes non résolus, d'« ambiguïtés », représentées par des étiquettes composites (*portmanteau tags*[27]), comme nom_verbe, pour l'hésitation entre nom et verbe. Pour un fragment de l'exemple de ***Mitterrand1*** fourni ci-dessus, un résultat non désambiguïsé serait :

{mot=je, lemme=je,		

corpus (Véronis et Khouri, 1995) insiste sur la nécessité de distinguer les descriptions lexicales, c'est-à-dire l'ensemble des associations trait-valeur qui caractérisent chaque forme, et les étiquettes, le passage des premières aux secondes se faisant par traduction, toute description lexicale devant correspondre à une étiquette au plus.

[27] Littéralement, des étiquettes-valises, sur le modèle de *portmanteau-word*.

catégorie=pronom}		
{mot=suis, lemme=être, catégorie=verbe}	{mot=suis, lemme=suivre, catégorie=verbe}	
{mot=la, lemme=le, catégorie=déterminant}	{mot=la, lemme=le, catégorie=pronom}	{mot=la, lemme=la, catégorie=nom}
{mot=France, lemme=France, catégorie=nom}		

où figurent les deux verbes correspondant potentiellement à la forme fléchie *suis* : *suivre* et *être*, tous deux légitimes hors contexte, et les trois étiquettes possibles pour *la*[28].

La degré d'étiquetage nécessaire à une expérience sur un corpus dépend étroitement des objectifs de la recherche envisagée. Si l'on veut se servir d'un corpus étiqueté pour extraire des suites de catégories syntaxiques, on peut tolérer un tel degré d'ambiguïté et trier *a posteriori* les résultats. Par contre, si l'on souhaite étudier un phénomène massif (comme la détermination) dans des gros corpus, on ne saurait se satisfaire d'un étiquetage qui laisse en suspens les choix (ici entre déterminant et pronom pour *le*, *la*, *les* ...).

8.4.3 Une vision large de l'étiquetage

Etiqueter un segment de texte (un mot, mais aussi un groupe de mots, une phrase, un paragraphe, etc.), c'est, de manière générale, lui associer des informations arbitrairement complexes[29]. Ces informations peuvent se situer à plusieurs niveaux de l'analyse linguistique : morphologie, syntaxe, sémantique, pragmatique , sans se limiter d'ailleurs aux aspects linguistiques (comme le trait diplôme utilisé pour ***Enfants*** ou le trait référence de ***Susanne***).

Cette vision élargie de l'étiquetage ne correspond cependant pas à l'acception la plus répandue. Quand on parle de corpus étiqueté, en particulier dans la communauté TALN, on fait référence le plus souvent à un document où chaque mot possède une étiquette morpho-syntaxique et une seule.

[28] Déterminant, pronom et nom (dans l'expression : *donner le la*).
[29] Nous avons fourni des structures de traits plates. Rien n'empêche d'employer des co-indiciations (Ligozat, 1994) assurant des partages de valeurs (on y a recours au chapitre suivant), ou encore des structures arbitrairement enchâssées qui regroupent des « paquets » de traits : DETMS est l'abréviation de {catégorie=déterminant, accord={genre=masculin, nombre=singulier}}, où le trait accord regroupe les traits de genre et nombre.

9. ÉTIQUETAGE PARTIEL ET TYPOLOGIE DE TEXTES

Le fait de disposer de textes partiellement étiquetés (un certain nombre de traits linguistiques fins sont privilégiés) permet d'entreprendre une typologie linguistique de ces textes, mais il n'est pas sûr qu'on puisse généraliser aisément les oppositions dégagées.

9.1 Circularité des démarches typologiques habituelles

La typologie des textes a suscité de nombreux travaux. Le plus souvent, ces recherches cherchent soit à caractériser les modes de production des textes (typologies situationnelles), soit à identifier les fonctions visées par les textes (typologies fonctionnelles). Les objectifs peuvent être didactiques (permettre à un apprenant d'identifier et de produire les différents types de textes de sa langue ou d'une langue étrangère) ou linguistiques, par exemple dans la lignée de la distinction histoire *versus* discours de Benveniste[30]. L'hypothèse partagée par ces différentes recherches est que chacun des types postulés se caractérise par l'association d'un certain nombre de caractéristiques linguistiques.

La démarche part souvent des types situationnels ou fonctionnels définis au départ, examine les textes qui relèvent de chacun de ces types et leur fonctionnement linguistique, et essaie de mettre en évidence certaines corrélations entre types et traits linguistiques. On ne sait toutefois pas si, en partant d'une autre typologie *a priori*, on ne rassemblerait pas sous un même chef des textes différents, ce qui aurait toutes chances de produire des agrégats de traits linguistiques distincts de ceux produits par la typologie précédente. La répartition des textes retenus sous les rubriques choisies est elle-même contestable. Il y a là une circularité d'autant plus gênante que l'existence de types textuels distincts paraît intuitivement fondée, même s'il s'avère délicat de l'étayer empiriquement.

9.2 Dégager les corrélations de traits linguistiques : D. Biber

Une autre optique consiste à faire émerger les types de textes grâce à un traitement statistique de textes étiquetés. C'est la ligne directrice des travaux de D. Biber (1988, 1989). Ce dernier examine les cooccurrences entre 67 traits linguistiques dans les 1 000 premiers mots[31] de 481 textes d'anglais contemporain écrit et oral. Ces textes proviennent de ***LOB*** et

[30] Elle oppose les énoncés reliés au moment de l'énonciation (emploi du présent, d'« embrayeurs » comme les pronoms de première et deuxième personne) : le *discours* à ceux qui effacent cet ancrage (emploi du passé simple, de la non-personne, c'est-à-dire la troisième personne) : l'*histoire*

[31] Cet échantillonnage a pour fonction de faciliter la comparaison des distributions de traits linguistiques. Cf. chapitre VII et chapitre IX.

London-Lund et relèvent de « genres » divers : articles de recherche, reportages, conversations, nouvelles radiophoniques ... Les traits étudiés ressortissent à 16 catégories distinctes comme marqueurs de temps et d'aspect, adverbes et locutions adverbiales de temps et de lieu, pronoms et pro-verbes, questions, passifs, modaux, coordination, négation... Ils sont identifiés automatiquement (en limitant au maximum la vérification manuelle)[32].

L'étiquetage mis en œuvre par Biber s'éloigne de l'étiquetage morpho-syntaxique pratiqué en général. Il est partiel et partial. Il est « inéquitable » : il s'intéresse à des fonctionnements linguistiques très spécifiques qu'il analyse en détail tandis qu'il en laisse d'autres dans l'ombre. Par exemple, il privilégie certains verbes (modaux) et certaines formes verbales (passif, présent ...), mais ne traite pas systématiquement l'ensemble des classes de verbes ni toutes les flexions verbales.

La statistique multidimensionnelle [33] est mise à contribution pour repérer les oppositions majeures entre associations de traits linguistiques. Elle rassemble les traits qui ont tendance à apparaître ensemble. Elle constitue dans le même temps les configurations de traits qui sont systématiquement évités par ces rassemblements. Cette démarche permet d'obtenir des pôles multiples, positifs et négatifs, correspondant à ces constellations positives et négatives. Ces pôles deux à deux constituent des dimensions. Chaque texte, par son emploi des traits linguistiques étudiés, se situe en un point déterminé de l'espace à *n* dimensions déterminé par cette analyse.

La typologie construite par D. Biber à partir des résultats de l'analyse factorielle s'organise autour de cinq dimensions. La première oppose les textes qui se caractérisent par l'usage de *do* comme pro-verbe, celui de *be* comme verbe principal, le présent, les démonstratifs, les contractions du type *don't*, la première et la deuxième personne du singulier, le pronom *it* aux textes qui favorisent les noms, les mots longs, des adjectifs attributs, les prépositions. Biber appelle cette première dimension *production impliquée* versus *production informationnelle*[34]. Les autres dimensions sont nommées l'*orientation narrative*[35] versus *non narrative*, la *référence dépendante*[36] ou non de la situation d'énonciation, la *visée persuasive apparente*[37] *ou non*, *le style impersonnel*[38] *ou non*. Biber souligne que les dimensions proposées à l'issue de l'interprétation des contrastes majeurs mis en évidence par l'analyse factorielle sont en fait des prototypes, des pôles de fonctionnements textuels. Chacune des dimensions mises en évidence oppose deux pôles, mais les textes concrets se situent en des points variés des « échelles » ainsi définies.

A partir de ces cinq dimensions, en utilisant des techniques de classification automatique[39], Biber aboutit à huit types de textes, en

[32] Ces traits et leur repérage sont décrits en détail dans (Biber, 1988, p. 211-245).
[33] Cf. chapitre IX.
[34] *Involved* versus *informational production*.
[35] Caractérisée par le passé, la 3e personne, la négation synthétique, les participes présents.
[36] Manifestée par les adverbiaux, en particulier de temps et de lieu.
[37] Les traits privilégiés comprennent les infinitifs, les modaux, les subordonnées conditionnelles.
[38] Favorisant les passifs sans agent et les passifs avec *by*.
[39] Cf. chapitre IX.

fonction de leur place sur chacune de ces dimensions :

1) Interaction interpersonnelle intime (*intimate interpersonal interaction*) ;

2) Interaction informationnelle (*informational interaction*) ;

3) Exposé « scientifique » (*« scientific » exposition*) ;

4) Exposé savant (*learned exposition*) ;

5) Fiction narrative (*imaginative fiction*) ;

6) Récit (*general narrative exposition*) ;

7) Reportage situé (*situated reportage*) ;

8) Argumentation impliquée (*involved persuasion*).

Ces types ne correspondent pas forcément aux intuitions communes. C'est ainsi qu'on ne débouche pas sur un type unique *interaction* ou *dialogue*, mais deux : l'interaction à visée informationnelle et l'interaction à visée interpersonnelle. De la même manière, Biber distingue plusieurs types de textes « expositifs » et de textes narratifs.

9.3 Généralité des typologies induites

Cette démarche permet la construction inductive d'une typologie de textes, basée sur les corrélations effectives entre traits linguistiques[40]. Elle court néanmoins le risque d'aboutir à des oppositions qui, pour avoir été établies à partir de textes concrets, ne valent que pour ces textes et pour les traits choisis pour les opposer[41]. Peut-on accorder une portée plus générale aux types ainsi construits ? Biber (1995) a appliqué la même démarche, mais cette fois-ci à quatre corpus, le corpus anglais initial et trois ensembles de textes en coréen, somali[42] et nukulaelae tuvaluan[43]. Malgré des différences nettes, liées en particulier au degré d'alphabétisation et à la place des traditions orales dans les langues considérées, Biber (*ibid.*, p. 359) pense pouvoir émettre l'hypothèse que les types textuels qu'il dégage sont communs à plusieurs langues, mêmes si leur réalisation linguistique diffère d'une langue à l'autre.

L'articulation de ces constats généraux, sur des corpus diversifiés, avec des analyses dans un domaine particulier ne va cependant pas de soi. Ainsi, Bergounioux *et coll.* (1982) étudient les résolutions générales votées par les congrés confédéraux des quatre centrales interprofessionnelles, CFDT, CFTC, CGT et FO, pendant les années 1971-1976. Ce corpus n'est pas étiqueté, soulignons-le. La répartition précise d'un certain nombre de formes (marques d'énonciation, détermination, coordination, pronoms, prépositions, etc.) dans les textes

[40] J.-P. Sueur (1982) étudie dans une optique très proche les contrastes entre parties de la Résolution Générale du congrès de 1976 de la CFDT. Il étiquette (manuellement cette fois) les traits qui lui paraissent pertinents et utilise là encore l'analyse factorielle des correspondances pour mettre en évidence les oppositions majeures.

[41] Il est intéressant à cet égard de comparer les traits retenus par Biber avec ceux choisis par Sueur (1982) et ceux privilégiés par Bronckart (1985).

[42] Langue parlée par environ 5 millions de personnes en Somalie, à Djibouti, en Éthiopie et au Kenya.

[43] Langue parlée par 350 personnes sur l'atoll Nukulaelae du groupe Tuvalu (Pacifique).

de ces quatre organisations syndicales a pour objectif de dégager l'organisation d'ensemble de ces textes (*ibid.,* p. 169-186). Un programme qui isole les mots qui sont significativement sur-employés dans une partie d'un corpus au regard de leur emploi dans le corpus entier[44] est utilisé pour évaluer les phénomènes étudiés. Ce programme dégage en même temps les sous-emplois significatifs d'une partie au regard du tout[45]. Les convergences des sur-emplois et des sous-emplois permettent d'opposer (*ibid.,* p. 175) une structure dite analytique, utilisée par la CFDT et la CGT à une structure dite déclarative, préférée par FO et la CFTC. Le premier type de résolution sur-emploie en particulier le verbe *être* à la troisième personne de l'indicatif présent, les modaux, les pronoms à la première personne du pluriel et les possessifs de même personne, les pronoms de troisième personne. Le deuxième type sur-emploie les verbes déclaratifs (*appelle*, *considère*, *estime*, *exige* ...), ayant pour sujet *le congrès* ou le sigle (*la CFTC*), suivis d'une complétive en *que*. Une autre étude (Habert, 1983), consacrée aux résolutions générales des congrès de la CFTC de 1945 à 1964 et de la CFDT de 1965 à 1979[46], trouve une opposition similaire. D'un côté une résolution « circonstancielle », ancrée dans le temps de l'énonciation : indications précises de lieu, verbes d'affirmation ou d'interpellation. De l'autre une résolution « théorique » qui s'affranchit de l'*ici et maintenant* de l'énonciation : présent de vérité générale (avec les flexions d'*être* et *avoir*), effacement de l'énonciateur, verbes modaux, marques d'articulation logique du discours, etc. Les résolutions examinées se situent entre ces deux pôles, la résolution « théorique » prenant le pas en 1945, moment d'affirmation du syndicalisme chrétien dans une France de l'après-guerre marquée par le rôle du Parti Communiste et de la CGT, et en 1970, 1973 et 1976 où la CFDT, après 1968, opte pour le *socialisme autogestionnaire*[47]. À travers ces deux études, l'une sur une période courte (5 ans), l'autre sur le moyen terme (34 ans), il semble que deux types de textes, au moins, soient disponibles pour permettre à un acteur social de se situer dans le présent, associés à des « postures » distinctes.

Les deux types de textes dégagés pour le discours syndical, très spécifiques, ne s'intègrent pas immédiatement dans ceux proposés par Biber, qui sont pourtant conçus pour rendre compte d'une grande diversité d'énoncés. La question de la généralité des typologies induites à partir des comportements observées reste donc encore largement ouverte.

[44] La présentation de la technique probabiliste correspondante est effectuée dans le chapitre IX.

[45] Soulignons deux apports de ce programme. La simple lecture ne perçoit qu'une partie limitée des sur-emplois effectifs. Elle est bien en peine de juger s'ils sont significatifs ou non. Les sous-emplois, le « creux » d'une partie au regard de l'ensemble, échappent le plus souvent à la conscience. Ils sont ici dégagés.

[46] La CFTC, centrale chrétienne, s'est transformée en 1964 en CFDT, une minorité constituant la CFTC « maintenue ».

[47] L'évolution récente de la CFDT vers plus de pragmatisme s'accompagne d'ailleurs d'une utilisation en congrès de formes proches de celles de la résolution circonstancielle.

10. ÉTIQUETAGE INTEGRAL ET SOCIO-STYLISTIQUE

Un étiquetage intégral bien que rudimentaire permet d'examiner les « parlures » d'un corpus regroupant des énoncés de plusieurs locuteurs de différentes catégories sociales.

10.1 Repérer les catégories et les suites de catégories de différents locuteurs

Enfants, une fois étiqueté et lemmatisé, a été étudié (Habert et Salem, 1995) sous l'angle de l'opposition entre locuteurs sans diplôme, titulaires du baccalauréat et personnes ayant suivi des études supérieures. Un des objectifs de l'utilisation d'une version étiquetée du corpus était de dégager des éléments caractéristiques des « parlures », des styles sociaux présents. Quelles sont les catégories morpho-syntaxiques privilégiées par chaque type de locuteur ? Quels sont les « patrons syntaxiques » qui leur sont propres ?

10.2 Varier le jeu d'étiquettes selon les phénomènes observés

Le corpus a été étiqueté par l'étiqueteur AlethCat[48]. Le « nettoyage manuel » qui a suivi a permis de rectifier un certain nombre d'erreurs de catégorisation et d'homogénéiser la lemmatisation des formes[49].

Les étiquettes employées par l'étiqueteur utilisé, une soixantaine au total, sont relativement rudimentaires : partie du discours, éventuellement sous-type dans la partie du discours (type de déterminant par exemple), traits morphologiques (verbe conjugué / infinitif / participe ..., genre, nombre, personne ...)[50]. Bien d'autres informations pourraient être associées aux « mots » : type de verbe (auxiliaire, modal ...), mot attendant des arguments...

A l'inverse, la présence de certaines indications (le genre, le nombre pour les noms et les adjectifs par exemple) peut rendre plus difficile la perception de certaines régularités : on disperse par exemple les occurrences de la catégorie des noms en masculin singulier, masculin pluriel, féminin singulier et féminin pluriel. Pour faciliter l'étude de telle ou telle opposition, on a donc transformé[51] le jeu d'étiquettes employé, soit en éliminant des informations présentes soit en en rajoutant.

---

[48] Développé par la société GSI-ERLI. Cet étiqueteur est conçu pour préparer le travail d'un analyseur syntaxique automatique.

[49] Notons que l'étiquetage automatique aboutit parfois à « souder » physiquement des constituants de « mots composés » (*bien_que*, *met_en_évidence*, *vis_à_vis_de* ...).

[50] Une étiquette spécifique non-réponse rend compte de l'absence d'une réponse à la question pour un locuteur donné.

[51] C'est un changement systématique ou à confirmer au coup par coup qu'on pourrait partiellement réaliser avec les fonctions de remplacement d'un simple traitement de textes.

Si l'on prend la phrase suivante :

> je ne sais pas, les gens sont égoïstes peut-être.

en faisant abstraction du lemme, après étiquetage et correction :

> <S01=31> je {PROPERS} ne {ADVNEG} sais {VIPR1S} pas {ADVNEG} , {PONCT-FAIBLE} les {DETDEF} gens {NOMMP} sont {VIPR3P} égoïstes {ADJMP} peut-être {ADV} . {PONCT-FORTE}

que l'on peut représenter aussi, pour plus de « clarté », de la manière suivante :

> <diplôme=études-supérieures, âge=-30>
>
> {forme=je, catégorie=pronom, type=personnel}
>
> {forme=ne, catégorie=adverbe, type=négation}
>
> {forme=sais, catégorie=verbe, mode=indicatif, temps=présent, nombre=singulier, personne=1}
>
> {forme=pas, catégorie=adverbe, type=négation}
>
> [...]

plusieurs transformations ont été utilisées :

- la réduction aux parties du discours traditionnelles :

> {diplôme=études-supérieures, âge=-30}
>
> {forme=je, catégorie=pronom}
>
> {forme=ne, catégorie=adverbe}
>
> {forme=sais, catégorie=verbe}
>
> {forme=pas, catégorie=adverbe}
>
> [...]

- l'élimination des marques de personne, genre et nombre pour les noms et les adjectifs :

> {diplôme=études-supérieures, âge=-30}
>
> [...]
>
> {forme=les, catégorie=déterminant, type=défini}
>
> {forme=gens, catégorie=nom}
>
> {forme=sont, catégorie=verbe, mode=indicatif, temps=présent}
>
> {forme=égoïstes, catégorie=adjectif}
>
> [...]

- l'ajout de la distinction entre adjectifs qualificatifs et adjectifs relationnels :
  Certains adjectifs sont en étroite correspondance avec des noms. Leur

étude complète donc celle de la répartition de cette catégorie majeure au sein du corpus. Ce sont les adjectifs relationnels. Rappelons leurs propriétés (Melis-Puchulu, 1991). Ce sont des adjectifs dénominaux : ils peuvent être mis en rapport avec des séquences *de* + nom comme dans *élection présidentielle* / *élection du président*. Ils ne sont pas gradables : **une carte très géographique*, et ne peuvent être employées de manière prédicative *: *cette carte est géographique*. Dans une séquence d'adjectifs post-posés, ils sont immédiatement après le nom, les adjectifs qualificatifs venant après *: une élection présidentielle surprenante* / **une élection surprenante présidentielle*. L'opposition n'est pas une opposition de nature, mais d'emploi. Ainsi, certains adjectifs relationnels ont également des emplois qualificatifs[52] *: *Cette politique est économique* / *Cette formule est très économique*.

Le résultat est ici :

{diplôme=études-supérieures, âge=-30}

[...]

{forme=les, catégorie=déterminant, type=défini}

{forme=gens, catégorie=nom}

{forme=sont, catégorie=verbe, mode=indicatif, temps=présent, nombre=pluriel, personne=3}

{forme=égoïstes, catégorie=adjectif, type=qualificatif}

[...]

Ces transformations, une fois effectuées, ont été soumises à l'analyse quantitative les différentes versions étiquetées du texte réduites à leurs seules étiquettes, ce qui donne pour l'étiquetage en parties du discours :

{diplôme=études-supérieures, âge=-30}

{catégorie=pronom}

{catégorie=adverbe}

{catégorie=verbe}

{catégorie=adverbe}

{catégorie=ponctuation}

[...]

ou encore en éliminant le nom du trait retenu :

{diplôme=études-supérieures, âge=-30}

{pronom}

{adverbe}

{verbe}

[52] Et inversement, certains adjectifs d'emploi surtout qualificatif peuvent se révéler relationnels selon le contexte. On trouve ainsi dans ***Menelas*** *syndrome douloureux thoracique*, où la place de *douloureux* entre le nom et un autre adjectif relationnel prouve que cet adjectif est ici relationnel. **Syndrome très douloureux* est d'ailleurs impossible dans ce domaine.

{adverbe}

{ponctuation}

[...]

Le programme d'analyse des sur-emplois et des sous-emplois évoqué *supra* permet d'opposer les locuteurs selon leur niveau d'études. Ce sont les étiquettes, pour chacun des jeux, qui sont soumises à examen, mais aussi les suites d'étiquettes, les segments répétés[53] constitués d'étiquettes. Une fois dégagées les tendances d'emploi des étiquettes et de leurs enchaînements, des outils de filtrage permettent d'extraire dans les textes catégorisés les séquences relevant des schémas syntaxiques retenus.

10.3 Une première opposition : style nominal et style verbal

L'examen des proportions relatives d'emploi des parties du discours selon les parties du corpus est instructive. La proportion des noms et des adjectifs croît avec le niveau de diplôme. À l'inverse, le domaine du verbal (verbes, adverbes, pronoms) décroît avec l'élévation du niveau d'études. Ce constat rejoint d'ailleurs ceux faits sur plusieurs corpus pour d'autres études socio-linguistiques. On notera la place éminente, toutes parties confondues, du nom (et des prépositions) : elle tient peut-être à ce que le type de question posée favorise des énoncés qui se présentent sous la forme d'un groupe nominal.

Si l'on s'en tient aux parties du discours seules et qu'on exclut les segments répétés dans lesquels elles entrent, les sans-diplômes se caractérisent par les non-réponses et par le sur-emploi du verbe (et des catégories associées : adverbe et pronom), les plus diplômés par le suremploi des adjectifs et de la coordination. Le faible nombre des catégories employées et le nombre important d'occurrences de chaque étiquette débouchent sur des segments répétés d'étiquettes extrêmement nombreux. On note la présence de syntagmes prépositionnels enchaînés chez les bacheliers comme : [{nom} {préposition} {déterminant} {nom} {préposition} {déterminant} {nom}], ainsi que le poids des adjectifs chez les diplômés du supérieur, en particulier dans des coordinations :

[{nom} {adjectif} {coordonnant} {adjectif}]

[{adjectif} {ponctuation}[54] {adjectif}]

[{nom} {adjectif} {ponctuation} {nom} {adjectif}]

[{déterminant} {nom} {adjectif} {ponctuation} {déterminant} {nom} {adjectif}]

La réduction du corpus aux seules parties du discours fournit une première approche de l'utilisation du matériel linguistique selon les types

[53] L'utilisation de segments répétés de formes ou d'étiquettes est présentée dans le chapitre IX.

[54] Ici comme dans les deux segments répétés suivants, il s'agit en fait de la virgule, dans son rôle de coordonnant.

de locuteurs. Certains phénomènes se trouvent cependant « écrasés » par cette réduction : le sur-emploi significatif de la catégorie adverbe chez les non-diplômés correspond dans près de la moitié des cas (354 occurrences sur 653) à des adverbes de négation. C'est sont alors les résultats obtenus avec un jeu d'étiquettes à mi-chemin du jeu restreint des parties du discours et de celui, trop éclaté, fourni par l'étiqueteur AlethCat qui ont été examinés. Il a semblé important de pouvoir disposer de sous-types des catégories « majeures » employées (à l'instar d'adverbe de négation par rapport à adverbe).

10.4 Examen des patrons syntaxiques caractéristiques de chaque type de locuteur

Cumulant des emplois multiples, les unités lexicales de la négation (*ne*, *pas*, *guère*, *jamais*, *que*) dominent les énoncés des sans-diplômes. Elles structurent le patron sur-employé [{pronom personnel} {adverbe négation} {verbe 1ère personne singulier} {adverbe négation}][55], de type *je ne vois pas*, *je ne sais pas*, ou le même patron suivi de la virgule : *je ne sais pas,* <reste de la réponse>, qui ne constitue pas exactement une réponse, mais une dévalorisation préalable de la réponse à venir. Par ailleurs, bon nombre d'exemples du patron [{adverbe négation} {verbe présent 3ème personne singulier} {adverbe négation}], sur-employé également par ces locuteurs, correspondent à l'indication par l'enquêteur de la difficulté à répondre chez la personne interrogée : *ne voit pas de raisons* (4 occurrences), *ne sait pas* (8 occurrences) et des variantes comme *ne peut pas répondre*. La non-réponse, comme silence (non-réponse), comme refus explicite de répondre, ou comme mise en doute préalable[56] des propos tenus, est centrale dans cette partie. Le patron sur-employé [{pronom personnel} {adverbe négation } {verbe 3ème personne singulier} {adverbe négation}] est dû pour l'essentiel à l'emploi négatif du présentatif *il y a* dans des réponses comme : *c'est la situation qui décide le logement si il n'y a pas de place* ou encore *quand il n'y a pas assez d'argent dans le ménage*. Il ne faut pas cependant en tirer des conséquences quant à l'orientation argumentative des réponses. Même si l'on trouve des séquences qui mentionnent des difficultés (*il n'y a pas de travail*, 3 occurrences dont l'une en contexte conditionnel), le présentatif négatif peut servir au locuteur à plaider au contraire pour le fait d'avoir des enfants. Les réponses suivantes en témoignent*: il n'y a pas de couple sans enfant* ou encore *il n'y a pas de raison valable*. Les sans- diplômes se caractérisent en outre par des phrases plus courtes, éventuellement réduites à un nom seul, non déterminé[57].

[55] Nous ne donnons que les valeurs des traits pour faciliter la lecture.
[56] Pour déterminer la place du phénomène, au début de chaque réponse a été introduit un « anti-point », noté {ponctuation début-phrase}. Les segments répétés comprenant cette étiquette confirment la tendance des locuteurs sans diplôme à commencer la phrase par un pronom personnel (en règle générale la première personne du singulier) suivi d'une négation : [{ponctuation début-phrase} {pronom personnel} {verbe indicatif présent 1ère personne singulier} {adverbe négation}{ponctuation faible}].
[57] Le motif [{nom} {ponctuation forte}] est sur-employé, les formes correspondantes les plus employées étant *chômage*, 11 occurrences, *égoïsme*, 8 occurrences et *argent*, 7 occurrences.

Les bacheliers sont caractérisés par les enchaînements de syntagmes prépositionnels, puisqu'on trouve des patrons comme : [{nom} {préposition} {article défini} {nom} {préposition} {article défini} {nom}] ou encore comme [{nom} {adjectif} {ponctuation faible} {nom} {préposition} {nom}]. Ce dernier patron est lié à des énumérations nominales, non déterminées (cf. l'absence de déterminant après la ponctuation faible) comme dans la réponse *: raison financière, situation de travail, peur de perdre son travail pour la femme qui s'absente pour raison de maternité*.

Les plus diplômés privilégient nettement l'adjectif et une forme qui en est proche, le participe passé, en particulier dans des coordinations, dans des patrons répétés comme [{nom} {adjectif} {coordonnant} {adjectif}] ou [{adjectif} {ponctuation faible} {adjectif}].

10.5 Préciser l'emploi des adjectifs : qualificatifs et relationnels

Hors contexte, les lemmes des adjectifs du corpus ont été répartis entre les catégories suivantes : {adjectif qualificatif} (*mauvais*), {adjectif relationnel} (*géographique*) et {adjectif qualificatif/relationnel} (*économique*). Cet enrichissement des étiquettes des adjectifs a ensuite été appliqué au texte : l'étiquette {adjectif} associée à *égoïstes* devient par exemple {adjectif qualificatif}. L'examen de la répartition des adjectifs relationnels par rapport aux qualificatifs permet de préciser le fonctionnement dans le corpus de la catégorie nominale prise au sens large .

Les adjectifs n'apparaissent pas dans les formes et segments sur-employés des non-diplômés. En ce qui concerne les bacheliers, seule la catégorie {adjectif relationnel} apparaît comme sur-employée, isolée ou dans des segments répétés. Ce sur-emploi souligne la nature nominale et prépositionnelle de cette partie (puisqu'un adjectif relationnel est équivalent à un syntagme prépositionnel). Cette équivalence est particulièrement flagrante dans le segment répété [{nom} {adjectif relationnel} {ponctuation faible} {nom} {préposition} {nom}] qui coordonne (par une virgule) un nom modifié par un adjectif relationnel et un nom dominant un syntagme prépositionnel. L'adjectif relationnel caractérise davantage les diplômés du supérieur. L'examen des contextes montre en effet que les adjectifs portant l'étiquette {adjectif qualificatif/relationnel} sont en fait tous relationnels dans cette partie : les constats quantitatifs sous-estiment donc la place des adjectifs relationnels. Voici quelques segments répétés significatifs :

[{déterminant défini} {nom} {adjectif relationnel}]

[{nom} {adjectif qualificatif/relationnel}]

10.6 Evaluation et perspectives

L'analyse des décomptes portant sur l'utilisation de divers jeux d'étiquettes donne une image intéressante de l'usage de l'appareil

linguistique par les différents ensembles de locuteurs : expression personnelle, modalisant la réponse faite, à dominante négative pour les sans-diplômes *versus* expression nominale, située hors du *ici et maintenant* pour les diplômés. Les bacheliers marquent une préférence pour les syntagmes prépositionnels, les diplômés du supérieur pour les adjectifs en particulier coordonnés. Les locuteurs ayant fait des études supérieures font appel plutôt aux adjectifs dénominaux qu'aux syntagmes prépositionnels pour modifier les noms, à l'inverse des locuteurs ne possédant que le baccalauréat. S'agirait-il d'un phénomène d'hypercorrection, d'une manière d'éviter le « style substantif » ?

Cependant, bien d'autres interprétations pourraient être produites pour les données constituées avec ces différents jeux d'étiquettes. Par exemple entre des réponses directes (baccalauréat et études supérieures) et des réponses différées (sans-diplômes), où les formules comme *je ne sais pas,* etc., ressemblent aux items de retardement de la réponse mis en évidence en analyse de la conversation.

11. UTILISER ETIQUETEURS ET CORPUS ETIQUETES

11.1 Adapter l'étiquetage aux objectifs de recherche

11.1.1 Un étiquetage est orienté par une famille de tâches

Meyer et Tenney parlent (1993, p. 25-26) d'étiquetage *finalisé (problem-oriented tagging*), à propos de l'étude de l'apposition dans *Survey of English Usage* faite par l'un d'eux. Ils ajoutent que les programmes d'étiquetage disponibles « sont moins utiles pour le linguiste travaillant sur corpus qui souhaite étudier une construction linguistique donnée en détail et adapter le jeu d'étiquettes qu'il met en œuvre pour étudier cette construction. »

Il faut généraliser ce constat. Un étiquetage est toujours orienté par une tâche, même si c'est implicite. Le jeu d'étiquettes utilisé permet d'étudier certains phénomènes ou de développer certains traitements ultérieurs, tandis qu'il laisse d'autres aspects linguistiques dans l'ombre et n'est pas compatible avec d'autres applications. Ainsi, la distinction du genre et du nombre pour les noms et adjectifs dans l'étiquetage d'***Enfants*** n'est pas forcément pertinente pour une étude énonciative de ce corpus, mais par contre, elle est utile pour une analyse syntaxique ultérieure : elle permet de vérifier des contraintes d'accord au sein du groupe nominal. À l'inverse, tous les étiqueteurs ne fournissent pas le temps et la personne pour les verbes conjugués[58], bien que cette information soit

[58] Ne serait-ce qu'en raison de la difficulté d'assurer une désambiguïsation efficace sur ce point : *travaille* est un présent de l'indicatif, 1ère et 3ème personne du singulier, mais aussi un présent du subjonctif aux mêmes personnes et enfin un impératif 2ème

particulièrement précieuse dans une perspective typologique comme celle de Biber. Une catégorisation donne ainsi à voir certains phénomènes et en ignore d'autres. Il faut donc multiplier les points de vue et à tout le moins être conscient des capacités heuristiques et des angles morts des jeux d'étiquettes auxquels on a recours. Les projets de comparaison et d'évaluation d'étiqueteurs se développent aujourd'hui (Paroubek *et al.*, 1997). Ce qu'on peut en attendre, ce n'est certainement pas une mise en évidence de la « meilleure catégorisation », ce qui n'a pas grand sens, mais l'identification des objectifs, points forts et faiblesses de chaque catégorisation et de l'adéquation de chacune aux projets de recherche envisagés.

11.1.2 Un étiquetage peut être « détourné »

Nous rencontrons avec les corpus étiquetés une situation courante pour les corpus annotés en général. L'annotation du corpus utilisé ne correspond pas exactement à la classification souhaitée des données, aux phénomènes que l'on souhaite isoler, au regard théorique que l'on porte sur eux. Pire : pour diverses raisons (le plus souvent le manque de moyens financiers et humains), il n'est pas possible de ré-étiqueter le corpus.

Il s'agit alors de « composer » avec l'état présent de l'étiquetage, d'en tirer les informations qui se rapprochent de celles recherchées. C'est cette démarche que nous avons vue à l'œuvre dans les études typologiques sur le discours syndical : faute de disposer de corpus étiquetés (il y a 15 ans, dans les limbes pour l'anglais et inexistants pour le français), on étudie aussi précisément que possible un ensemble délimité de formes graphiques (de « mots »), malgré le " bruit " introduit par l'utilisation de cette représentation sommaire.

A l'inverse, si, dans le cas présent, une telle démarche typologique peut se satisfaire, pour un premier dégrossissage, de corpus « bruts », c'est-à-dire réduits à des formes graphiques, elle gagne sans conteste à utiliser des corpus étiquetés de manière spécifique. L'écart entre les données utilisées par ces différentes analyses et la plus ou moins grande immédiateté d'interprétation qui en résulte débouche néanmoins sur la nécessité plus générale de vérifier l'adéquation possible (au prix de détournements éventuels) entre les conventions d'annotation du corpus utilisé et les objectifs de recherche visés.

11.1.3 Le ré-étiquetage est incontournable

L'écart entre les catégories associées à un corpus déjà catégorisé ou fournies par un étiqueteur accessible et celles dont on peut avoir besoin pour une étude donnée implique souvent une recatégorisation (partielle) du corpus. Nous avons montré comment, pour ***Enfants***, l'ajout d'une

---

personne du singulier.

nouvelle distinction (adjectif qualificatif / relationnel) venait préciser une étiquette existante. Le ré-étiquetage peut aussi conduire à des révisions plus drastiques, lorsque les choix de segmentation de départ sont remis en cause (le choix des « mots composés » pertinents pour le corpus en cause) ou quand certains phénomènes sont traités différemment (par exemple, *rapide* analysé tantôt comme un adjectif tantôt comme un adverbe dans *Prenons une rapide décision*).

Le ré-étiquetage total ou partiel peut aussi avoir comme visée l'alignement des résultats de deux étiqueteurs sur un même corpus, à des fins de comparaison ou d'évaluation (Atwell *et al.*, 1994). Selon Belmore (1994, p. 52) : « Une manière d'utiliser les corpus pour améliorer de manière cumulative les analyses consiste à déterminer les différences exactes entre deux analyses d'un même corpus. Dans l'idéal, l'une des deux analyses partirait de la première et représenterait alors un essai explicite d'amélioration. »

11.2 Environnements de catégorisation et de manipulation de texte étiqueté

Paradoxalement, il semble que le besoin d'environnements informatiques de catégorisation et de manipulation de texte étiqueté, souvent souligné par les participants des projets d'étiquetage et de structuration de corpus, reçoive dans l'immédiat peu de réalisations concrètes (Greenbaum et Yibin, 1994, p. 44).

11.2.1 Catégoriser

On peut vouloir étiqueter, totalement ou partiellement, un texte « nu ». S'il s'agit d'utiliser un corpus déjà étiqueté ou les résultats d'un étiqueteur disponible, la finesse des distinguos nécessaires pour des analyses proprement linguistiques suppose des programmes permettant de préciser l'étiquetage morpho-syntaxique accompagnant désormais nombre de corpus. Elle implique aussi des modules de catégorisation interactive ou de modification interactive d'étiquetages préalables, certaines valeurs d'étiquettes ne pouvant pas être attribuées automatiquement[59].

11.2.2 Manipuler des corpus étiquetés

Les programmes nécessaires ici permettent d'extraire du texte étiqueté des motifs arbitrairement complexes. Les constituants de ces motifs sont,

[59] Par exemple, la distinction entre déterminants définis spécifiques *versus* génériques dans (Sueur, 1982).

ici encore, des structures de traits[60]. Le motif (ou patron) correspondra au fragment de texte pour lequel les structures de traits de ses composants s'apparient avec celles des éléments correspondants du texte. On parle de filtrage (*pattern-matching*). Des opérateurs permettent la conjonction, la disjonction, l'optionalité, la répétition de ces contraintes, etc. Par exemple, le motif :

> [{nom} {adjectif relationnel ∨ qualificatif/relationnel} {coordonnant} {adjectif relationnel ∨ qualificatif/relationnel}]

permet de chercher les noms suivis de deux adjectifs coordonnés soit relationnels soit qualificatifs ou relationnels (c'est ce qu'indique la disjonction ∨).

De tels environnements facilitent le nécessaire retour au contexte qui permet d'éviter les commentaires oiseux de simples artefacts. Dans ***Enfants***, par exemple, les réponses fournies par les plus diplômés paraissent plus riches en séquences du type : [{adjectif qualificatif} {nom}]. Ce résultat attire l'attention : en français moderne, l'antéposition de l'adjectif est une construction de langue tenue. Déception : l'examen des séquences relevant de ce patron montre qu'en fait, il s'agit souvent d'adjectifs modifiant un nom antéposé. L'ambiguïté est due à l'absence de marque de ponctuation entre les groupes nominaux dans des suites de formes comme : *temps libre argent*.

12. ENJEUX THEORIQUES

12.1 Le dit est le dire

L'examen des catégories employées, et des segments de catégories conduit à s'attacher aux patrons syntaxiques des énoncés, voire aux genres textuels qui peuvent expliquer le recours à tel type de construction. D'autres phénomènes linguistiques s'offrent à une exploration méthodique et à la quantification. Nous espérons avoir montré qu'une analyse du dire (du style, du mode de parler) était tout aussi instructive qu'une analyse du dit. Le détour par des catégories abstraites, ici morpho-syntaxiques, introduit une bienfaisante étrangeté dans l'appréhension du corpus. Ce pas de côté contrebalance la trompeuse immédiateté des formes lexicales, dont le sens s'impose trop évidemment. Mais, en même temps, certaines associations de traits dans les dimensions dégagées par Biber ou le sur-emploi de telle étiquette dans l'étude d'***Enfants*** demeurent énigmatiques. On ne dispose pas forcément dans l'immédiat des cadres théoriques nécessaires pour examiner les données ainsi produites.

[60] Là encore, comme en 1.3, nous fournissons une représentation unifiée des différentes possibilités effectives dans tel ou tel système d'interrogation de texte étiqueté.

12.2 Linguistique et textualité

Benveniste assignait à la linguistique la phrase comme horizon d'analyse. Il n'en a pas moins exploré les régularités proprement textuelles liées à l'utilisation de l'appareil de l'énonciation. Sa distinction histoire / discours a donné naissance à des typologies ou des grilles d'analyse plus fines. En didactique du français, cette dichotomie a été mobilisée largement pour aider les apprenants à maîtriser les conditions de bonne formation des textes.

L'utilisation de corpus étiquetés diversifiés offre désormais la possibilité d'examiner sérieusement l'hypothèse que les textes effectifs relèvent de types fondamentaux qui expliquent un certain nombre de leurs traits linguistiques. Les études existantes en fournissent des caractérisations empiriques fines. La généralité des catégories dégagées, leur lien aux genres et registres intuitivement distingués par les locuteurs restent à travailler. Ces résultats appellent peut-être un renouveau de la linguistique textuelle : on attend un modèle de la compétence textuelle qui intègre les contraintes détaillées mises en évidence.

Il reste également à explorer le fonctionnement social des types de textes disponibles dans une communauté langagière donnée. Pour Bakhtine : « Nous ne parlons qu'à travers certains genres discursifs, c'est-à-dire que tous nos énoncés possèdent certaines formes relativement stables et typiques pour se constituer en totalités » (Todorov, 1981, p. 129). L'organisation de chacun de ces genres est socialement significative : « Le genre forme [...] un système modélisant qui propose un simulacre du monde » (*ibid.,* p. 128). C'est le cas des deux types de résolutions de congrès syndicaux évoqués en 2.3. La résolution déclarative (ou circonstancielle) va de pair avec un refus de tenir un discours global sur la société. Elle légifère essentiellement pour le laps de temps qui la sépare du congrès suivant. La résolution analytique (ou théorique) s'installe dans l'éternel présent de la théorie, dépassant les limites du *ici et maintenant*. L'idéologie s'y exprime dans de longs développements sans locuteur explicite. Les articulations logiques veulent entraîner dans un réseau d'enchaînements qui font appel au simple examen de la « nature des choses »[61].

12.3 Analyses multi-dimensionnelles

Toutes proportions gardées, les études typologiques à partir de formes graphiques seules ou à partir de traits linguistiques clairement identifiés (cf. 2) tiennent de la reconstitution d'animaux disparus à partir de fossiles épars et incomplets. Au vu de données langagières fragmentaires, on

[61] Il est intéressant à cet égard de noter que l'utilisation par la CFTC et la CFDT de ces deux types de résolutions ne se superpose pas mécaniquement à l'évolution historique de cette confédération. La déconfessionnalisation de 1964 n'entraîne pas un changement sur ce plan. C'est quand cette organisation veut affirmer fortement un projet social propre qu'elle recourt à la résolution théorique : en 1945 — dans l'immédiat après-guerre, et après 1968.

postule l'existence d'un squelette syntaxique[62] voire textuel. On fait l'hypothèse de dépendances fonctionnelles entre des éléments relevant de niveaux distincts de l'analyse linguistique. Avec le risque d'inventer des « monstres langagiers » sans existence réelle.

Les techniques d'analyses statistiques multi-dimensionnelles comme l'analyse factorielle des correspondances utilisée par Biber ont précisément pour objectif de manifester les corrélations effectives entre des variables multiples. Elles mettent en évidence des régularités qui échappent à l'observation « à l'œil nu ». Elles débouchent sur des regroupements de comportements langagiers qui peuvent renouveler nos analyses des dépendances entre niveaux linguistiques[63]. Elles manifestent des oppositions qui restructurent notre catégorisation préalable des données.

[62] Comme la proto-phrase donnée comme sous-jacente aux résolutions déclaratives (Bergounioux *et al.*, 1982, p. 178) évoquées en 2.3 : « considérant [...] le congrès [...] {verbe déclaratif 3ème personne présent} [...] que [...] {subjonctif} [...] {déterminant indéfini}. »

[63] Pour poursuivre la métaphore, notons que l'apport de ces méthodes a été considérable en classification des espèces : elles ont permis d'améliorer les taxonomies existantes, limitées dans leur capacité à percevoir et organiser des corrélations multiples.

CHAPITRE II

LES CORPUS ARBORES

Nous montrons dans une première section les notations employées pour rendre compte des relations syntaxiques et nous rappelons la nature des phénomènes à noter. Nous présentons dans une deuxième section un corpus arboré, ***Susanne***, qui représente une réalisation exemplaire par la finesse de l'annotation produite et par la manière dont les choix effectués sont documentés. La troisième section est consacrée à l'utilisation de corpus arborés et de parseurs pour l'étude de la phraséologie. La dernière section examine les enjeux théoriques de corpus arborés et les conditions pratiques de leur emploi.

13. DIVERSITE DES CORPUS ARBORES

Même si les jeux d'étiquettes varient et si les séquences à catégoriser, selon qu'on regroupe ou non des unités polylexicales, il est relativement aisé de se faire une idée d'un corpus étiqueté : une catégorie est associée à chaque occurrence du texte. Cette belle simplicité disparaît dès qu'on aborde les corpus arborés, c'est-à-dire « décorés » d'arbres. L'annotation résultante peut varier du tout au tout. Il s'agit en effet de délimiter des groupes, de les nommer (les catégoriser), et de statuer sur leurs relations. À ces trois niveaux, les points de vue sont multiples. Tous les constituants ne font pas l'unanimité : c'est le cas du syntagme verbal, cher à la tradition chomskyenne, et rejeté par M. Gross. G. Sampson (Sampson, 1995, p. 4) cite à ce propos une expérience significative. À la rencontre de 1991 de l'Association for Computional Linguistics, des chercheurs en TALN appartenant à neuf institutions différentes se sont vu demander de délimiter les constituants d'un ensemble de phrases. Pour l'exemple suivant, voici les seuls parenthésages qui ont fait l'unanimité :

He said this constituted a [very serious] misuse [of the [Criminal Court] processes].

Nous définissons les principales facettes des corpus arborés : les notations disponibles, la manière d'obtenir les analyses, les types d'analyses et d'analyseurs, les niveaux d'annotation syntaxique.

13.1 Noter des relations syntaxiques

13.1.1 Arbres, graphes et relations

Les arbres sont le dispositif habituel pour noter les relations syntaxiques. La tradition veut que les feuilles soient à la base et la racine au sommet. On distingue les *nœuds terminaux*, les feuilles et les autres nœuds, appelés *non-terminaux*. Ces nœuds non-terminaux englobent les *nœuds pré-terminaux*, qui dominent directement les feuilles. Si l'on considère un nœud et ses fils, un arbre matérialise deux relations particulières : celle de dépendance immédiate, entre le père et ses fils, et celle de précédence immédiate, entre un nœud aîné et son ou ses cadet(s).

A un nœud est associée une étiquette (SN) et éventuellement des « décorations » : une série d'associations trait-valeur du type {genre=masculin, nombre=singulier}. Comme pour l'étiquetage morpho-syntaxique, les étiquettes simples ou complexes se ramènent en fait toutes à une structure de traits : SN = {cat=SN} et SNMS = {cat=SN, genre=masculin, nombre=singulier}. Conceptuellement, à chaque nœud, correspond donc non une étiquette mais une structure de traits.

Les deux relations de dépendance et de précédence ne suffisent pas à noter la variété des phénomènes syntaxiques. Deux nœuds frères séparés par d'autres nœuds peuvent constituer une unité discontinue (comme la négation complète *ne ... pas* dans *Ne me quitte pas*). L'anaphore suppose un lien entre l'anaphorique et son antécédent : il s'agit d'un lien entre des nœuds qui ne sont généralement pas frères, mais à des niveaux différents de la représentation syntaxique. Certains constituants sont « flottants » : leur insertion à un endroit donné ne rend pas compte de leur portée réelle. C'est le cas des adverbes de phrase comme heureusement dans *Heureusement Jean a terminé son année* et *Jean a heureusement terminé son année*. L'attachement réel d'un nœud peut rester en suspens, même pour un locuteur : c'est le cas dans *Jean a heureusement terminé son année*, où *heureusement* peut modifier la phrase dans son ensemble, mais également le syntagme verbal seul (*il est heureux que Jean ait terminé son année* / *Jean a terminé son année d'une manière heureuse*).

Visiblement l'arbre ne suffit plus à noter tous ces phénomènes. On peut souhaiter recourir à des graphes moins limités, où un nœud peut être le point d'arrivée de plusieurs arêtes. Il faudrait même que ces graphes puissent être « polychromes »[64] pour visualiser aisément les diverses

[64] Cf. (Marandin et Cori, 1993) pour une proposition formelle en ce sens.

relations à l'œuvre. Une autre direction de travail consiste à utiliser des descriptions logiques d'arbres, où l'on ne manipule ni des arbres ni des graphes, mais la conjonction logique des divers types de relations identifiées entre les nœuds. Elle est explorée par Vijay-Shanker (1992), dans la lignée des travaux de M. Marcus (Marcus *et al.*, 1983).

Cette remise en cause de l'arbre comme mode fondamental de notation syntaxique n'est pas nouvelle[65]. Elle peut plus profondément renvoyer au choix entre grammaires de constituants et grammaires de dépendances.

13.1.2 Grammaires de constituants et grammaires de dépendance

On trouve dans Tesnière les prolégomènes des grammaires de dépendance. I. Mel'cuk, qui s'inscrit dans cette lignée, contraste (1988, p. 12-42) les grammaires de dépendance avec les grammaires de constituants (*phrase structure grammars*). Les grammaires de constituants mettent au premier plan l'inclusion d'un segment dans une catégorie syntagmatique et des segments d'un type dans des segments de niveau supérieur (deux constituants sont ou bien enchâssés ou bien disjoints). La plupart des nœuds y sont non-terminaux. Les nœuds d'un niveau donné sont ordonnés linéairement. Les relations de domination sont entre constituants et non pas entre mots. Les grammaires de dépendance révèlent les liens hiérarchiques entre mots. Tous les nœuds sont terminaux. Ils ne suivent pas forcément un ordre linéaire. Un arbre de dépendance du type [V sont reçus [N [N Pierre][Coord et][N Jacques]] ne contient aucune information directe concernant l'ordre linéaire des mots dans l'énoncé, qui peut se réaliser sous la forme *Pierre et Jacques sont reçus* comme sous la forme *Sont reçus Pierre et Jacques*.

Ce sont les grammaires de constituants qui sont majoritairement employées pour les corpus annotés syntaxiquement. La langue traitée peut expliquer le choix fait. Les grammaires de constituants semblent mieux adaptées aux langues à ordre des mots relativement contraint et aux syntagmes nettement identifiables, comme l'anglais[66]. Les grammaires de dépendance conviennent davantage aux langues où l'ordre des mots est plus libre (le finnois, par exemple). Contribuent sans doute également à cette prépondérance le poids des travaux proprement linguistiques qui relèvent de cette tradition mais aussi le fait que la technologie des parseurs pour les langages informatiques fait aussi appel aux grammaires hors contexte. Les grammaires de dépendance offrent cependant l'avantage de faciliter l'utilisation des relations hiérarchiques entre mots d'un énoncé. Si l'on veut dégager les cadres de sous-catégorisation des verbes, par exemple, cette approche permet un

[65] On trouve dans le modèle GPSG (Gazdar *et al.*, 1985) la volonté de découpler dans les règles hors contexte la relation de dominance et l'ordre linéaire, c'est-à-dire la précédence.

[66] Toutefois, le parseur ENCG - English Constraint Grammar (Karlsson *et al.*, 1995), crée des structures de dépendance pour l'anglais. (Karlsson, 1994, p. 130-142) fournit plusieurs exemples de résultats commentés (extraits d'un manuel informatique, d'*Alice au pays des merveilles* et d'une encyclopédie). Inversement, certains formalismes cherchent à rendre compte des variations d'ordre des mots dans le cadre des grammaires de constituants.

élagage immédiat qui ne conserve que les liens de dépendance pertinents.

13.1.3 Notations textuelles

Puisque les arbres constituent la notation prépondérante, nous continuons à parler de corpus arborés. Le stockage d'arbres pour leur traitement informatique suppose de passer d'une représentation dans le plan à une représentation textuelle essentiellement linéaire : elle figure par l'enchâssement la relation de dépendance et par la succession la relation de précédence. Des dispositifs annexes permettent de dépasser les limites des arbres. Il s'agit généralement d'indices attachés aux nœuds et de renvois à ces indices pour exprimer les autres relations.

Le format de présentation des corpus arborés varie. Il peut être horizontal : c'est le cas de cet exemple[67] emprunté à la banque d'arbres d'IBM France :

> [N Ce_DEDEMMS guide_NCOMS N][V [P leur_PPCA6MP P] permet V_VINIP3 [P de_PREPD [Vi se_PPRE6MP familiariser_VPRN [P avec_PREP [N les_DARDFP opérations_NCOFP [P de_PREPD [N réseau_NCOMS [A local_AJQMS A] N] P][A effectuées_VTRPSFP [P par_PREP [N les_DARDMP utilisateurs_NCOMP N] P] A] N] P] Vi] P] V] ._.

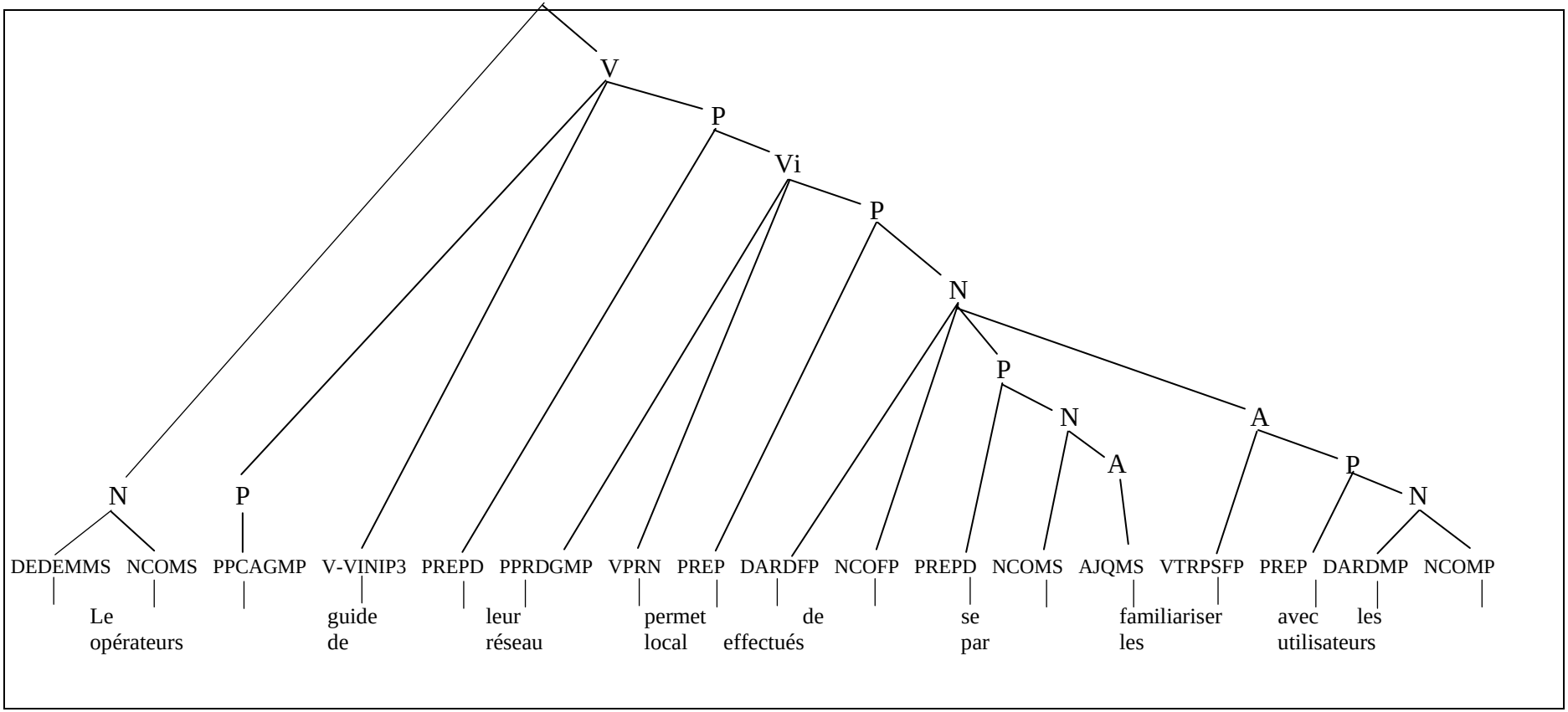
L'étiquette du constituant est souvent fournie deux fois : au début et à la fin du groupe en cause, probablement pour faciliter le repérage visuel des groupes et des frontières. Les enchâssements font apparaître une hiérarchie, dont l'indentation, plaçant les constituants de même niveau à une même distance de la marge gauche, facilite la perception :

```
[N Ce_DEDEMMS guide_NCOMS N]
[V
 [P leur_PPCA6MP P]
    permet V_VINIP3
    [P de_PREPD [Vi se_PPRE6MP familiariser_VPRN [...]
```

Il peut également être vertical. On distingue comme dans ***Susanne*** formes, étiquettes de mots, parties d'arbres. Pour l'exemple choisi :

[67] Cité dans (Leech *et al.*, 1996, p. 6).

Ce	DEDEMMS	[N .
guide	NCOMS	. N]
leur	PPCA6MP	[V [P . P]
permet.	V_VINIP3	.
de	PREPD	[P .
se	PPRE6MP	[Vi .
familiariser	VPRN	.
avec	PREP	[P .
les	DARDFP	[N .
opérations	NCOFP	.
de	PREPD	[P .
réseau	NCOMS	[N .
local	AJQMS	[A . A] N] P]
effectuées	VTRPSFP	[A .
par	PREP	[P .
les	DARDMP	[N .
utilisateurs	NCOMP	. N] P] A] N] P] Vi] P] V]
._.		


V
P
Vi
P
N
P
N
A
P
N
A
N
DEDEMMS NCOMS PPCAGMP V-VINIP3 PREPD PPRDGMP VPRN PREP DARDFP NCOFP PREPD NCOMS AJQMS VTRPSFP PREP DARDMP NCOMP
Le opérateurs
guide de
leur réseau
permet local
de effectués
se par
familiariser les
avec utilisateurs
les

Le mot figure en première colonne, sa catégorie en seconde. La troisième colonne fournit une partie de l'arbre syntaxique : le point y marque l'insertion du sous-groupe constitué de la catégorie et du mot. Les deux premières lignes correspondent ainsi au sous-arbre [N [DDEMMS Ce][NCOMS guide] N].

Ces deux présentations, verticale et horizontale, correspondent à l'arbre donné dans la figure ci-contre[68] (nous le simplifions en omettant les catégories pré-terminales).

13.2 Obtenir des analyses

Il est possible d'associer à un texte des annotations syntaxiques plus ou moins complexes de manière purement manuelle. Mais, sauf à disposer de moyens humains et matériels très importants, cela limite la taille du texte ainsi analysé. C'est le choix qui a été fait pour ***Susanne*** (cf. section 2), parce qu'il s'agissait d'obtenir une analyse aussi fouillée que possible. C'est encore le cas des corpus qui sont balisés à la main pour servir de corpus d'apprentissage de grammaires probabilistes (cf. chapitre VIII), comme celui développé en commun par l'université de Lancaster et IBM (Eyes et Leech, 1993). Dans ce cas (*ibid.* p. 132), à l'opposé de ***Susanne***, il s'agit d'insérer des arbres dits « squelettiques » (parenthésage et catégorisation des constituants principaux). L'autre possibilité est l'analyse syntaxique automatique, ou parsage, mieux adaptée au traitement de gros volumes textuels.

Entre « travail manuel » et parsage, bien des intermédiaires existent : l'intervention humaine peut se produire en amont (pour délimiter des groupes ou éliminer des catégories parasites) ou en aval pour trancher entre plusieurs analyses : c'est le cas du système TOSCA (Halteren et Oostdijk, 1993, p. 154) ou pour améliorer l'analyse produite : c'est la solution retenue par ***Penn Treebank*** (Marcus *et al.,* 1993).

13.3 Types d'analyse

13.3.1 Analyse partielle / analyse complète

L'analyse peut être partielle ou complète. Complète : c'est un arbre qui couvre l'ensemble de la phrase, dont les feuilles sont les mots de la phrase. Partielle : à une phrase donnée correspond(ent) un ou plusieurs arbre(s) qui laisse(nt) des parties qui ne sont pas analysées.

Une analyse partielle peut correspondre à l'incapacité du parseur, pour une phrase particulière ou en général, à produire des structures qui couvrent l'intégralité des données analysées. Mais une analyse partielle

[68] *A priori*, il est toujours possible de passer automatiquement d'un format à un autre, et d'en fournir une version réellement arborée comme ici, même si le détail du codage propre à tel corpus peut rendre difficile la mise au point du traitement nécessaire.

peut correspondre aussi au fait de ne s'intéresser qu'aux composants d'une certaine nature syntaxique. C'est ainsi qu'en terminologie automatisée, les extracteurs de groupes nominaux se concentrent sur ces syntagmes, où figurent les dénominations polylexicales du domaine. Dans la phrase suivante de ***Mitterrand1***[69] : « le Louvre , libéré du le[70] ministère des les finances , cela représente un immense palais , le plus grand musée du le monde — un kilomètre sept cent si vous voulez en faire le tour — imaginez la fatigue des les pieds des les visiteurs : il faut que les œuvres d' art soient quand même à la portée de ceux qui veulent se déplacer », sont retenus par LEXTER (cf. 3.4), à partir de la version lemmatisée, les groupes nominaux suivants :

[SN [SAdj [Adj immense]][SN [Nom palais]]]

[SN [SAdj [Adj grand]][SN [SN [Nom musée]][SP [Prep de][SN [Det [Art le]][SN [Nom monde]]]]]]

[SN [SN [Nom fatigue]][SP [Prep de][SN [Det [Art le]][SN [SN [Nom pied]][SP [Prep de][SN [Det [Art le]][SN [Nom visiteur]]]]]]]

[SN [SN [Nom œuvre]][SP [Prep de][SN [/Nom art]]]]

Une analyse partielle peut enfin avoir pour but de produire une version simplifiée de la phrase, en laissant de côté des composants ou des parties de composants conçus comme secondaires. Par exemple, le parseur peut extraire l'association sujet – verbe – complément d'objet, et ignorer les compléments circonstanciels, si l'objectif est d'étudier la sous-catégorisation des verbes, leurs cadres syntaxiques et leurs arguments typiques.

13.3.2 Une seule analyse ou plusieurs

Le résultat peut fournir, pour un segment donné, une seule analyse ou plusieurs.

On distingue deux types d'ambiguïtés. Ambiguïtés réelles : un locuteur ne pourrait pas trancher. Hors contexte, par exemple, il est difficile de savoir comment analyser *état de l'art abstrait* (*Cette thèse commence par un [état de l'art] abstrait* / *Ce critique d'art présente l'état de l'[art abstrait]*). Ambiguïtés techniques : le savoir dont dispose le parseur n'est pas suffisant pour choisir entre des possibles[71], mais un locuteur n'a pas de difficultés à le faire, en fonction de ses connaissances générales ou au vu du contexte[72]. C'est le cas des rattachements prépositionnels et

[69] Emission de TF1 *Ça nous intéresse, Monsieur le Président*, du 28 avril 1985.
[70] Dans le pré-traitement, les contractions préposition + article défini (*aux*, *du*, *des*) sont décomposées pour faciliter les opérations ultérieures.
[71] T. Briscoe (1994, p. 99) donne l'exemple de la définition de *youth hostel (A hostel for usu. young people walking around country areas on holiday for which they pay small amounts of money to the youth hostels association or to the international yha)* dans le *Longman Dictionary of Contemporary English* (*LDOCE*). Le parseur inclus dans Alvey Natural Language Tools, avec un dictionnaire de 20 000 entrées, a produit plus de 2 500 analyses. Voir (Souter et Atwell, 1994, p. 151) pour un autre exemple d'analyse ambiguë.
[72] À l'inverse, un annotateur confronté à des phrases isolées peut se trouver dans l'incapacité de trancher (Black *et al.*, 1993, p. 40).

adjectivaux. Dans l'expression *traitement du langage naturel*, s'il ne dispose pas dans son lexique de l'expression *langage naturel*, un analyseur peut ne pas savoir s'il faut rattacher *naturel* à *traitement* ou à *langage*.

Voici, à titre d'exemple, les pourcentages d'ambiguïté obtenus par le système TOSCA sur un corpus d'1,5 million de mots de prose anglaise contemporaine (Halteren et Oostdijk, 1993, p. 155) :

Nombre d'analyses différentes	fiction	non-fiction
1	22 %	20 %
2	15 %	15 %
3-5	17 %	19 %
6-10	15 %	15 %
11-20	10 %	12 %
21-100	15 %	16 %
> 100	6 %	3 %

Ces chiffres donnent une idée des difficultés rencontrées en analyse syntaxique automatique.

13.3.3 Sous-spécification

Il est possible de laisser une analyse sous-spécifiée, c'est-à-dire incomplète sur un point donné. Cela revient à limiter artificiellement l'ambiguïté, en la laissant implicite. Par exemple, les attachements prépositionnels ou adjectivaux, souvent difficiles à effectuer automatiquement, peuvent être " laissés en suspens " pour permettre une post-édition spécifique. Le parseur ENGCG (Voutilainen et Heikkila, 1994, p. 190) dans *fat butcher's wife*, indique juste que *fat* s'attache à un nom à droite sans décider s'il s'agit de *butcher* (*la femme du gros boucher*) ou de *wife* (*la grosse femme du boucher*) et n'effectue pas non plus les rattachements des adverbiaux, notoirement délicats. C'est encore le cas du parseur Fidditch (Hindle, 1994) dans ***Penn Treebank*** qui ne rattache pas les groupes dont il ne peut pas déterminer avec certitude le rôle dans une structure de plus haut niveau (cf. chapitre VIII). Cela peut aboutir à fournir pour une phrase une suite d'arbres non reliés entre eux. Dans certains cas, des nœuds sont laissés sans étiquette quand leur délimitation est claire, mais pas leur catégorie (Black *et al.,* 1993, p. 19).

13.4 Analyseurs de texte « tout-venant »

Nous précisons les types de parseurs qui sont effectivement employés pour l'annotation de vastes corpus, ainsi que les choix qui conditionnent leur fonctionnement : production d'une seule analyse ou de plusieurs,

analyse descendante ou montante.

Certains formalismes syntaxiques contemporains comme LFG, HPSG, les grammaires d'arbres adjoints (Abeillé, 1993) ou comme le modèle Gouvernement et Liage ont donné lieu à la réalisation de parseurs. Toutefois, ces analyseurs sont avant tout destinés à tester le traitement par ces formalismes de phénomènes linguistiques complexes (dépendances à distance, etc.). S'ils visent à avoir la « couverture » la plus large possible, il faut entendre cet objectif comme la capacité à traiter un à un la plupart des problèmes syntaxiques d'une langue et non comme la capacité à traiter l'enchevêtrement de ces problèmes dans des phrases authentiques longues et complexes, qui peuvent même violer certaines « règles » grammaticales. Les parseurs de ces obédiences ne semblent pas dans l'immédiat utilisables sur de vastes corpus[73]. À notre connaissance, il n'existe d'ailleurs pas de corpus annoté selon leurs principes. Par opposition aux parseurs avant tout destinés à tester des formalismes syntaxiques raffinés, l'objectif des analyseurs qui sont évoqués dans ce chapitre est le parsage robuste. Il s'agit, pour reprendre les critères[74] de F. Karlsson (1994, p. 122), de pouvoir analyser, sans se bloquer, du texte « tout-venant », (en fournissant éventuellement des résultats partiels), d'aboutir à un taux satisfaisant d'analyses correctes[75] (*i.e.* où les mots sont dominés par une étiquette syntaxique unique et adéquate) et de ne pas aboutir à des résultats aberrants pour des phrases de longueur et de complexité « raisonnable ». D. Hindle (1994, p. 105) rejoint cette caractérisation. Il insiste en outre sur le fait que le parseur doit toujours produire « quelque chose », même sur un énoncé non grammatical. Il tient, mais c'est un point qui ne fait pas l'unanimité, à ce qu'un résultat et un seul soit retourné pour une phrase donnée. Il souhaite enfin que le parseur permette une amélioration incrémentale.

Les langages artificiels (langages de programmation, langages de représentation de connaissances) sont conçus *a priori* pour éviter toute ambiguïté : quand un programme est exécuté, son comportement, à un moment donné de son exécution, avec des données déterminées, doit être univoque. L'ambiguïté est au contraire centrale pour les langues naturelles. Elle est souvent ressentie comme une difficulté pour les traitements automatiques. Beaucoup de parseurs pour les langues naturelles ont pour visée la production de l'ensemble des analyses possibles. Ce peut être le cas au niveau de la phrase dans son ensemble, comme dans le système TOSCA. Ce peut être aussi le cas en analyse partielle. Certains analyseurs, en revanche, visent à ne fournir qu'une seule analyse. C'est le cas de Fidditch (Hindle, 1994), utilisé pour ***Penn Treebank***. Cette deuxième possibilité, à l'évidence, facilite la production de gros volumes de texte arboré, puisque le post-traitement manuel n'a pas à trier parmi les possibles.

L'objectif d'une ou de plusieurs analyses complètes pour du texte tout-

[73] Certains chercheurs pensent même que ces modèles avant tout théoriques sont de peu de profit pour développer des analyseurs utilisables, au contraire des grandes grammaires descriptives (Black *et al.*, 1993, p. 77).

[74] Nous ne reprenons pas son exigence de rapidité, pour des raisons expliquées au chapitre VIII.

[75] F. Karlsson (*ibid.*) cite l'objectif, qui paraît extrêmement ambitieux de 90 % d'analyses justes. Cf. les pourcentages d'ambiguïté fournis en 1.3.2.

venant est encore loin d'être réalisable. Les parseurs capables de produire des résultats partiels sont donc nécessaires, ce qui favorise les analyseurs montants. Les analyseurs montants (*bottom-up*) regroupent progressivement des structures de niveau de plus en plus élevé, les analyseurs descendants (*top-down*) suivent une approche inverse : des niveaux supérieurs vers les mots. Les premiers sont plus appropriés que les seconds pour fournir des résultats partiels : en quelque sorte, ils « savent » s'arrêter en chemin, en produisant des groupes qui ne sont pas forcément tous reliés, mais qui peuvent déjà être utilisés.

13.5 Niveaux d'analyse

L'examen des corpus arborés existants permet dans (Leech *et al.*, 1996, p. 9) de distinguer, par ordre de complexité croissante, les niveaux d'annotation suivants[76], illustrés sur l'exemple utilisé *supra* :

13.5.1.1 Simple parenthésage des constituants

Ce sont en fait des crochets qui sont le plus souvent utilisés :

[Ce guide][[leur] permet [de [se familiariser [...]

13.5.1.2 Étiquetage des constituants

C'est la représentation fournie plus haut (dans cet exemple, seules les étiquettes des nœuds pré-terminaux sont plus complexes).

On appelle *parsage squelettique* (*skeleton parsing*) le fait de s'en tenir à ces deux niveaux, voire au premier seul. Ce « dégrossissage syntaxique », qui peut être effectué manuellement à relativement faible coût, peut suffire à certaines analyses automatiques ultérieures (recherche de cadres de sous-catégorisation) ou servir de base d'entraînement à un analyseur probabiliste (cf. chapitre VIII).

13.5.1.3 Indication des relations de dépendance

Elle fournit les liens entre les « gouverneurs » (Tesnière ou Mel'cuk) ou « têtes » et leurs « dépendants »[77]. Leur notation se fait par des flèches. Ces liens relient uniquement des mots, à la différence des grammaires de constituants, où les ensembles reliés peuvent correspondre aussi bien à des mots qu'à des groupes de mots.

Nous empruntons les notations du parseur ENGCG (Voutilainen et Heikkila, 1994) pour illustrer cette approche sur notre exemple (> indique que la tête est à droite, la première des deux catégories suivant l'arrobas,

[76] D'autres informations sont distinguées pour un corpus d'oral transcrit et les caractéristiques syntaxiques propres à l'oral : réitérations, faux démarrages, etc. Nous ne les présentons pas, puisque nous avons fait le choix de ne traiter que les corpus d'écrit.

[77] Nous suivons ici la terminologie de (Mel'cuk, 1988, p. 23). La dénomination *dépendant* y est préférée à celle de *modifieur*, parce qu'elle est plus générique.

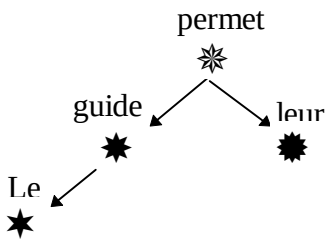
@, renvoie au mot examiné, la seconde au mot tête) :

Ce	@DN>
guide	@NV2>
leur	@PV>
permet	
[...]	

@DN> signifie que *Ce* est un Déterminant dépendant du premier Nom à droite (si c'était le deuxième, la notation serait @DN2>). Une autre notation, indiquée dans (Leech *et al.*, 1996, p. 26) assortit chaque mot d'un numéro d'ordre à sa gauche et éventuellement à droite du numéro de la tête dont il dépend :

1	Ce	D	2
2	guide	N	4
3	leur	P	4
4	permet	V	
[...]			

ce qui correspond à :



Le mot 1 (*Ce*) dépend du mot 2, qui, comme le mot 3, dépend du mot 4. Ce dernier, qui est la « tête », ne dépend de rien. Il est encore possible (*ibid.*, p. 27) de représenter un graphe de dépendance par une expression parenthésée où chaque parenthèse ouvrante est suivie d'une tête, puis des dépendants de celle-ci, et ce de manière récursive[78] :

[V permet [N guide [D Ce]][P leur][...]

13.5.1.4 Indication des relations fonctionnelles

Il s'agit de noter les fonctions comme sujet, objet direct, objet indirect etc. :

[N <Sujet> Ce_DEDEMMS guide_NCOMS N][V [P <ObjetIndirect> leur_PPCA6MP P] permet V_VINIP3 [...] ._.

13.5.1.5 Classification plus fine des syntagmes

Elle peut être assurée par un système de traits : [N{genre=masc, nombre=sing} Ce_DEDEMMS guide_NCOMS N][V{mode=indicatif, temps=présent, personne=3} [P{nombre=plur} leur_PPCA6MP P] permet V_VINIP3 [...] ._.

13.5.1.6 Relations " logiques " ou profondes

Il s'agit d'indiquer les liens de co-référence, de rassembler les constituants discontinus. Dans le cas présent, un indice (entre chevrons) peut manifester la coréférence entre leur et le sujet implicite (explicité par

[78] Du moins dans les cas où il n'y a pas de discontinuités.

un constituant vide) de se familiariser :

> [N Ce_DEDEMMS guide_NCOMS N][V [P <8> leur_PPCA6MP P] permet V_VINIP3 [P de_PREPD [N <8> N] [Vi se_PPRE6MP familiariser_VPRN [...] ._.

Ces constituants vides peuvent servir ensuite à faciliter le repérage des relations prédicat / arguments dans les phrases (Marcus *et al.,* 1993, p. 321).

13.5.1.7 Information sur le rang d'une unité syntaxique

Le niveau d'enchâssement des constituants est ajouté (il peut le plus souvent être calculé en fonction du niveau de parenthésage).

14. UNE REALISATION EXEMPLAIRE : *SUSANNE*

Susanne est un sous-ensemble de ***Brown*** qui avait déjà été manuellement analysé à Gothenburg. Il comprend 64 extraits de 2 000 mots chacun, soit 128 000 mots, relevant de quatre des genres distingués par ***Brown*** : reportage journalistique, Belles Lettres, écrit scientifique et technique, aventure et fiction. Le corpus obéit à un format vertical, comme nous l'avons vu au chapitre précédent, avec un mot par ligne, et dans l'ordre la référence, le statut (correction ou non), la catégorie pré-terminale, le mot, son lemme, et l'analyse syntaxique.

14.1 Une annotation « exhaustive »

Nous choisissons de présenter en détail ce corpus arboré « manuellement » pour trois raisons.

En premier lieu, c'est l'un des plus faciles d'accès, gratuitement et sans formalités.

En second lieu, le schéma d'annotation est l'un des plus documentés qui soit (Sampson, 1995) : les choix faits sont discutés en détail, ils sont exposés dans des documents aisément accessibles. Cela permet de comprendre et d'utiliser pleinement le résultat : « Les conventions d'annotation de ***Susanne*** proposent une méthode pour représenter tous les aspects de la grammaire anglaise qui sont suffisamment définis pour être susceptibles d'une annotation formelle. Les catégories et les limites entre elles sont spécifiées de manière suffisamment détaillée pour que, dans l'idéal, deux analystes annotant indépendamment le même texte et se référant aux mêmes conventions soient forcés de produire la même analyse structurale » (Sampson, 1994, p. 169).

Enfin, ***Susanne***, comme le souligne l'acronyme : Surface and Underlying Structural ANalyses of Natural English, vise une annotation

aussi exhaustive que possible (pratiquement tous les niveaux définis *supra* y sont représentés) (*ibid.* p. 170) : « son but (comparable à celui de la taxonomie de Linné au dix-huitième siècle dans le domaine de la botanique) n'est pas d'identifier les catégories qui sont optimales sur le plan théorique ou qui reflètent nécessairement l'organisation psychologique de la compétence linguistique des locuteurs, mais simplement d'offrir un schéma de catégories et des façons de les utiliser qui rende aisé aux chercheurs en TALN l'enregistrement systématique et sans ambiguïté de l'usage réel, sans malentendus sur des emplois locaux d'une terminologie analytique*. »* En ce sens, ***Susanne***, qui résulte d'une annotation entièrement humaine, explore les limites de l'annotation syntaxique. Nombre des annotations que ce corpus fournit ne pourraient pas être ajoutées automatiquement à d'autres corpus, au moins dans l'immédiat. En disposer, de façon expérimentale et sur un corpus de taille réduite, permet cependant d'évaluer l'intérêt de chacune d'entre elles pour les recherches, tant linguistiques que computationnelles.

14.2 Informations fournies dans Susanne

Voici les choix faits pour ***Susanne*** aux différents niveaux d'analyse définis *supra*.

352 étiquettes sont utilisées pour l'étiquetage des mots. Sampson (1995) fournit pour les catégories fermées la liste exhaustive et pour les catégories ouvertes les critères d'attribution. Les noms propres sont répartis en noms de personne, noms de lieux etc.

Les nœuds portent jusqu'à trois types d'information : catégorie, fonction et indice (permettant de relier le nœud à un autre nœud).

Les relations fonctionnelles suivantes sont indiquées : sujet logique, objet direct logique, objet indirect logique, agent du passif, sujet de surface, objet de surface, circonstants de lieu, de direction, de temps, de manière etc.

Les étiquettes catégorielles fournissent de nombreuses informations sur les constituants ainsi nommés (forme et type de verbe pour les groupes verbaux par exemple).

Des indices lient les paires de nœuds pour montrer l'identité référentielle entre des constituants qui se trouvent dans certaines configurations syntaxiques. Une étiquette spécifique dans le champ réservé au mot représente la « trace *»* : c'est-à-dire la position logique d'un constituant placé en fait ailleurs ou qui est effacé dans la structure syntaxique de surface. Simultanément, un constituant « déplacé *»* porte une autre étiquette marquant ce déplacement et un indice le lie à la « trace *»* correspondant à sa position « logique *».* Dans l'exemple suivant[79], *John wanted to go* :

[Nns:s123 John] wanted [Ti:o s123 to go]

[79] (Leech *et al.*, 1995, p. 19)

:s indique la fonction sujet, :o la fonction objet de l'infinitive (Ti) pour le verbe *wanted*. Le « fantôme » s123 indique la position logique du sujet de surface *John*. L'indice 123 établit le lien entre la réalisation de surface et le « fantôme ».

Les conventions de notation des étiquettes des nœuds permettent de distinguer les étiquettes pré-terminales, celles des syntagmes, celles des propositions et celles des unités « racines ».

15. PHRASEOLOGIE ET TRAITEMENTS SYNTAXIQUES

Les corpus arborés sont disponibles depuis le début des années quatre-vingt dix, c'est-à-dire depuis moins longtemps que les corpus étiquetés, accessibles depuis les années quatre-vingts. La primauté de l'anglais se fait ici écrasante : il n'existe pas à ce jour de corpus arboré du français aisément disponible[80].

En TALN, ces corpus servent surtout à la mise au point des parseurs. L'observation de corpus arborés permet de préciser les règles à employer, d'analyser automatiquement des corpus de taille plus importante, de retravailler les règles en jeu et ainsi de suite. Cette utilisation est évoquée au chapitre VIII. Les corpus arborés servent également d'étapes vers des traitements sémantiques (cooccurrences syntaxiques et similarités). Le chapitre IV traite cet aspect.

Les recherches linguistiques qui ont recours à des corpus arborés sont donc encore rares. Nous centrons notre analyse sur le traitement de la dimension phraséologique du langage, pour la langue générale — ce sont les « expressions figées », les « mots composés » mais surtout en langage de spécialité — ce sont les termes. C'est une zone à la lisière de la syntaxe et du lexique (Corbin, 1992). Nous présentons des utilisations de corpus arborés et d'analyseurs « robustes » pour rendre compte, en français et en anglais, de ces fonctionnements langagiers.

15.1 Le renouveau des études linguistiques de la phraséologie

Les expressions toutes faites, comme les noms composés (*un champignon atomique*), les verbes composés (dans des constructions à verbe support comme *mettre en évidence*), les locutions adverbiales (*à la volée*), prépositionnelles (*à la fin de*) ou conjonctives (*à seule fin que*), ont souvent été reléguées aux marges des traitements lexicographiques[81]. D'abord, ces unités polylexicales s'insèrent malaisément dans les

[80] Le Centre Scientifique d'IBM France a cependant développé au début des années quatre-vingt dix un corpus arboré de 400 000 mots (débats en français du parlement canadien, manuels IBM) qui peut être acheté. Nous en donnons un exemple *infra*.

[81] En français du moins. Les dictionnaires d'expressions idiomatiques foisonnent pour l'anglais.

dictionnaires sur support papier[82]. Où faire figurer *champignon atomique*, sous l'entrée *champignon* ou sous *atomique* ? Le rattachement à *champignon* paraît naturel, toutefois, c'est bien d'énergie nucléaire qu'il s'agit, et on souhaiterait maintenir ce lien. Où faire entrer *à la volée* ? Ces locutions sont d'ailleurs soumises à déformation (la réalisation originelle *goulet d'étranglement* est concurrencée par *goulot d'étranglement*), mais si les dictionnaires déconseillent certaines variantes, ils ne répertorient pas pour autant toutes les variantes effectives. Ensuite, on voit souvent dans ces séquences la partie « imagée *»,* métaphorique de la langue, comme le souligne A. Rey (Rey et Chantreau, 1979, p. I-XIII), ce qui conduit alors à privilégier une étude de l'origine et de l'évolution de ces séquences et peut-être à sous-estimer leur place dans la langue courante : « un dictionnaire de locutions, s'il n'est pas un simple recueil de traductions, ne peut être qu'historique *»* (*ibid.,* p. XII). Enfin, les limites de l'ensemble considéré sont floues, et variables les critères qui permettent de dire qu'une séquence fonctionne comme un « mot composé ». Si l'on considère *verre à vin* comme un nom composé, faut-il en faire de même de toutes les séquences similaires : *verre à cognac*, *verre à apéritif*, *verre à kyr* ... ?

La maîtrise de ces « mots en plusieurs mots *»* est pourtant essentielle dans l'apprentissage d'une langue. Ils s'avèrent en effet souvent opaques dans la phase de compréhension et causes d'hésitations dans la phase de production. C'est pourquoi Mel'cuk leur donne une place centrale dans son *Dictionnaire Explicatif et Combinatoire du Français*. Ses fonctions lexicales (Mel'cuk, 1988) visent à mettre au jour les réalisations lexicales les plus probables des mots pour exprimer une modification sémantique donnée. Le degré fort se dit ainsi *à chaudes larmes* quand il s'agit de *pleurer* et *à tout rompre* quand le verbe est *applaudir*.

Depuis une quinzaine d'années, la phraséologie suscite un renouveau d'intérêt en linguistique ainsi qu'en TALN. Dans la lignée logique des études menées sur les possibilités combinatoires des mots simples, qui soulignaient les multiples restrictions existantes (Guillet, 1990), les études du LADL ont montré l'importance des « mots composés *».* Elles ont abouti en particulier à un dictionnaire électronique des mots composés en français (Silberztein, 1993). Ce dictionnaire constitue un inventaire extrêmement poussé des expressions, sur le plan quantitatif, mais aussi sur le plan qualitatif. Chaque entrée est assortie de la description de ses variantes possibles. En TALN, l'évolution des formalismes vers la lexicalisation, c'est-à-dire la réduction des règles « générales *»* au profit de règles rendant compte des particularités d'emploi des mots sinon un par un, du moins par classes réduites, s'est accompagnée d'un renouveau des études et des propositions de traitement des expressions dites figées[83].

L'étude des unités polylexicales a conduit un certain nombre d'auteurs (Gazdar *et al.,* 1985 ; Abeillé, 1993 ; Habert et Jacquemin 1995) à postuler que ces unités relèvent des règles générales de la grammaire, mais

[82] Il n'en va bien sûr pas de même pour un dictionnaire électronique. Les fonctions de recherche permettent de séparer l'entrée concernée et les points d'accès.
[83] Cf. (Abeillé, 1993) pour une présentation sur ce point dans trois formalismes contemporains.

qu'elles obéissent à des contraintes supplémentaires[84], et qu'en particulier elles sont moins flexibles que les syntagmes libres de même catégorie : par exemple, on ne peut dire en conservant le même sens *#champignon très atomique*[85] ou *#champignon atomique et dangereux*, etc. Dans la logique de cette approche, on peut examiner une séquence qui constitue éventuellement une unité polylexicale, étudier les transformations syntaxiques dont elle est passible, et en tirer un constat global sur le « degré de figement » de cette séquence. L'hypothèse est que, plus une séquence est figée, c'est-à-dire moins elle accepte de transformations syntaxiques, plus il y a de chances qu'il s'agisse d'une unité polylexicale. C'est l'hypothèse défendue par G. Gross (1988).

L'apport des corpus à ce double renouveau porte sur deux points. En premier lieu, étant donné une expression jugée « contrainte » quant à ses possibilités de transformation, les corpus permettent de chercher si ses réalisations effectives confirment ce jugement. C'est ce que nous examinons en 3.2 et en 3.3 pour des expressions de la langue générale et des termes techniques, respectivement. Deuxièmement, l'ensemble des unités polylexicales est par définition ouvert. C'est par ce biais notamment que s'enrichit le lexique, en particulier dans les domaines techniques et scientifiques. L'observation des corpus sert alors à accroître le lexique des expressions. C'est ce que nous montrons pour les langages de spécialité en 3.4.

15.2 La flexibilité en corpus d'expressions polylexicales

H. Barkema (1993, 1994) se fixe pour objectif la « mesure » de la flexibilité réelle, en corpus, d'expressions toutes faites. Il examine donc les variations, c'est-à-dire les suites de mots qui sont apparentées à ces expressions et qui résultent d'une transformation graphique, phonétique, morphologique ou syntaxique (*gagner le cocotier* pour *gagner le coquetier* résulte d'une approximation phonétique, par exemple). Certaines de ces variations constituent des variantes, c'est-à-dire des équivalents effectifs de l'expression en cause (*infarctus myocardique* pour *infarctus du myocarde*, par exemple).

15.2.1 Les variations en corpus d'expressions « toutes faites »

Pour effectuer le repérage de telles variations, Barkema (1994) recherche les occurrences d'expressions courantes et les suites de mots qui en sont proches dans un vaste corpus, celui de Birmingham, qui rassemble 20 millions de mots. Ce corpus fournit par exemple 111 occurrences

[84] (Barkema, 1993) s'inscrit dans la même vision de hiérarchies de contraintes, tout comme, dans un autre cadre (van der Linden, 1992).

[85] Comme dans (Gazdar *et al.*, 1985) et (Barkema, 1994, p. 42, note 8), le # signale que la séquence en cause est grammaticale mais qu'elle ne peut pas être interprétée « idiomatiquement ». Elle pourrait dénoter un champignon fortement irradié et ne peut pas renvoyer au nuage caractéristique d'une explosion atomique.

inchangées de l'expression *cold war*[86] (*guerre froide*) ainsi que les 13 exemples suivants qui en constituent des variations :

1	renewed Cold War
2	the melting Cold War
3	the world Cold War
4	continuing, ever-present 'cold' war
5	the Cold War won by Europeans who 'destalinized' Eastern Europe
6	the cold war which threatened to divide the world into two ideological armed camps
7	a not-so-cold war against Kaddafi
8	the awkward cold war thought up by the American paranoids, who should be back in the law offices of middlewestern towns
9	a period of cold and hot civil war which ended with Hitler's invasion of Austria
10	a kind of cold civil war
11	the cold war that existed between the two giants, the United States and ...
12	the Cold War in Washington
13	the cold war between the Nature Conservancy Council and the farmers

Barkema répartit variations et emplois non modifiés selon le schéma syntaxique auquel ils obéissent :

Schéma	occurrences et numéros
[{déterminant} cold war]	111 occ.
[{déterminant} {adjectif} cold war]	3 occ. (1, 2, 4)
[{déterminant} cold war {proposition}]	2 occ. (6, 11)
[{déterminant} cold war {syntagme prépositionnel}	2 occ. (12, 13)
[{déterminant} cold war {participe passé}]	1 occ. (5)
[{déterminant} {adjectif} cold war {participe passé}]	1 occ. (8)
[{déterminant} Adv cold war {syntagme prépositionnel}]	1 occ. (7)
[{déterminant} {nom} cold war}	1 occ. (3)
[{déterminant} cold {adjectif} war}	1 occ. (10)
[{déterminant} cold {coordonnant} {adjectif} {adjectif} war {proposition}]	1 occ. (9)

15.2.2 " Mesurer " la flexibilité

Après cette première étape de recueil, Barkema se fixe pour objectif d'évaluer, et même de « mesurer » la flexibilité observée. Les variations effectives de la séquence dans un corpus jugé représentatif sont-elles prévisibles ? Au contraire, sont-elles plus importantes ou moins

[86] L'étude précise de cette séquence s'inscrit dans une recherche plus vaste : l'examen des variations de 450 expressions dans le même corpus (Barkema, 1993).

importantes que ce à quoi on pouvait s'attendre ?

L'hypothèse sous-jacente est que la flexibilité dépend au premier chef du schéma syntaxique de départ de la séquence examinée. Pour pouvoir porter un jugement sur ces variantes observées, c'est-à-dire déterminer si *cold war* est aussi flexible qu'on pourrait s'y attendre, il faut d'abord caractériser la flexibilité effective du schéma sous jacent : [{adjectif} {nom}].

Barkema utilise alors le corpus de Nimègue (130 000 mots), entièrement arboré et qui contient 16 183 syntagmes nominaux relevant de à 1 736 patrons syntaxiques distincts. Il compte le nombre d'occurrences du schéma [{adjectif} {nom}], avec un adjectif « absolu » et un {nom commun singulier} ainsi que le nombre d'occurrences des variantes syntaxiques de ce schéma (dont le passage au pluriel). Il compare alors la fréquence obtenue pour une variation de *cold war* relevant d'un patron donné avec la fréquence attendue. La fréquence attendue d'une telle variation s'obtient en multipliant le nombre total d'occurrences de *cold war* et de ses variations par le nombre de fois où le patron de cette variation se réalise dans les syntagmes libres[87] par rapport au nombre d'occurrences du schéma dont relève *cold war* et de ses variations au sein des syntagmes libres.

Dans les 16 183 syntagmes nominaux du corpus de Nimègue, 1 257 relèvent du schéma [{adjectif absolu} {nom commun singulier}], et 3 171 de ce schéma et de ses variantes syntaxiques. On s'attendrait alors à trouver 49,15 occurrences du schéma de base ((111 + 13) x (1 257 / 3 171)), alors qu'on en trouve 111 : la réalisation au singulier *cold war* est notablement plus fréquente que prévu, ce qui signifie aussi que *cold war* présente moins de variations que le schéma syntaxique dont elle relève ne le permet. L'examen des écarts entre les fréquences attendues et les fréquences observées souligne le fait que la post-modification de *cold war* par un syntagme prépositionnel est moins fréquente qu'on ne s'y attendrait. Il en va de même de la réalisation au pluriel (0 rencontrée, 24,64 occurrences attendues).

15.2.3 Évaluation

L'approche de Barkema pourrait être améliorée. Dans l'idéal, il faudrait pouvoir opérer sur le corpus de Birmingham qui a servi à extraire les variantes de *cold war*. Malheureusement, ce vaste corpus n'est pas muni de structures syntaxiques. Comme Barkema le souligne lui-même, il faudrait pouvoir calculer le poids de chaque réalisation syntaxique d'un schéma fondamental sur le même corpus que celui utilisé pour extraire les variations d'expressions relevant de ce schéma. En effet, rien ne dit que la flexibilité des syntagmes libres ou celle des expressions toutes faites soit la même dans tous les registres. On sait par exemple que l'écrit journalistique contemporain français fait souvent appel à des locutions qui sont détournées : par exemple ce titre de *Libération* du 20 mars 1989 après les élections municipales *Coup d'état de grâce* (Fiala et Habert,

[87] C'est-à-dire ne constituant pas des expressions toutes faites.

1989, p. 91). D'autres registres, comme le discours juridique, sont peut-être plus conservateurs quant à la phraséologie qu'ils véhiculent. Ne disposant pas de corpus arboré de taille suffisante pour pouvoir y observer des phénomènes de flexibilité, Barkema, par la force des choses, en est réduit à « peser » les variations effectives avec une balance réglée sur d'autres données langagières, le corpus de Nimègue, ce qui constitue un biais dont on ne peut pas mesurer les conséquences dans l'immédiat.

Barkema cherche à caractériser la flexibilité du schéma de base dont relève une expression donnée. Une partie des recherches actuelles en syntaxe met l'accent sur les contraintes lexicales gouvernant l'application des règles syntaxiques. Tout adjectif par exemple n'accepte pas la totalité des règles de formation des groupes adjectivaux ni ne rentre dans toutes les places syntaxiques possibles (antéposé / post-posé / après copule). Nous avons vu au chapitre I les restrictions propres aux adjectifs relationnels : construction copulative et adverbe de degré sont impossibles. Les adjectifs de couleur présentent d'autres particularités. Barkema examine simplement les variations du patron [{adjectif absolu} {nom commun singulier}]. C'est sans doute une caractérisation encore trop grossière[88]. Cependant, s'il paraît nécessaire d'utiliser des catégories plus fines, c'est accroître en amont la difficulté de disposer d'un corpus à la fois suffisamment vaste et étiqueté avec suffisamment de finesse.

15.3 La variation de termes en langue de spécialité

Pour obtenir les variations possibles de *cold war*, Barkema utilise un programme qui cherche les phrases comprenant *war* au singulier ou au pluriel et *cold*, pas forcément conjoints ni dans cet ordre. Le tri des séquences effectivement pertinentes est par contre manuel. Dans certaines d'entre elles, *cold* et *war* n'appartiennent pas au même syntagme ou bien ne suivent pas la relation de dépendance présente dans l'expression source.

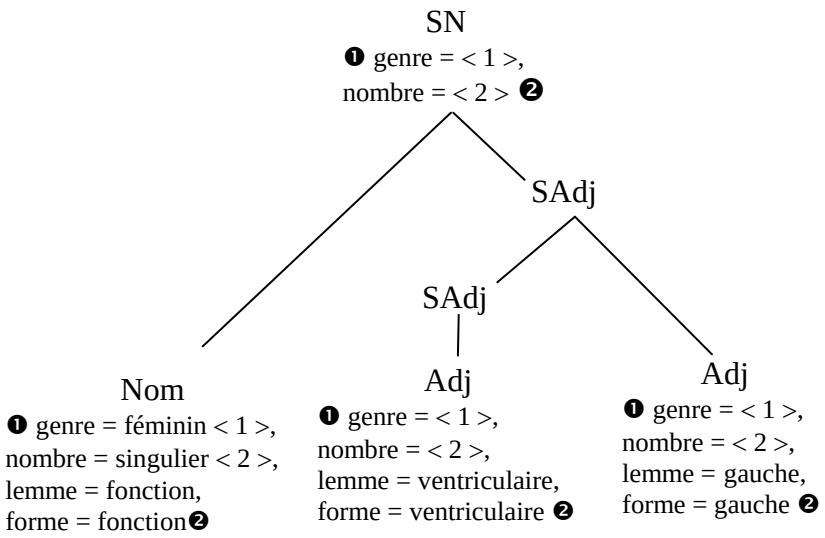
Les recherches de C. Jacquemin (Jacquemin, 1994) sur la variation des termes en langue de spécialité empruntent une démarche radicalement différente où la quête de variations est contrôlée par des connaissances, des règles linguistiques. Au lieu de chercher des séquences en intersection — c'est-à-dire partageant des mots — avec des expressions toutes faites, il s'agit d'engendrer les variations syntaxiques possibles de termes techniques et de vérifier si ces variations se rencontrent effectivement en corpus.

[88] Bien qu'il postule que : « [...] en principe, les expressions libres acceptent l'application de toutes les règles (et sont donc totalement flexibles) » (*ibid.*, p. 44), Barkema montre d'ailleurs quelque inquiétude sur ce point et souhaite vérifier pour des expressions libres comme *the old man ou the bird in the garden* si les variations effectives de ces expressions correspondent bien au profil de variations attendues.

15.3.1 Une représentation syntaxique contrainte des termes

L'objectif est d'inventorier les variations en corpus des termes d'un domaine. On parle aussi de mots-clés ou de *descripteurs* quand ces éléments sont utilisés en informatique documentaire pour indexer des documents. Certains de ces descripteurs sont des mots simples (comme *paradigme* en linguistique). La plupart sont des mots complexes (comme *axe paradigmatique* en linguistique). Ce sont les descripteurs complexes qui sont retenus.

Dans l'optique retenue par Jacquemin, les termes complexes ne sont pas représentés comme des simples suites de mots, mais directement comme des arbres syntaxiques aussi profonds et aussi larges que souhaité. Les relations de dépendance entre les composants sont donc directement indiquées. En outre, les nœuds de ces arbres sont décorés de traits également aussi complexes que nécessaire. Ces nœuds permettent d'assortir les arbres de fines contraintes de bonne formation. Ainsi, pour ***Menelas***, le descripteur *fonction ventriculaire gauche*[89] est représenté de la manière suivante[90] :



La représentation choisie souligne la dépendance de *gauche* par rapport à *ventriculaire* et non à *fonction*. On constate par ailleurs que le nombre de *fonction* est spécifié : ce doit être le singulier, si bien que la séquence *fonctions ventriculaires gauches* ne saurait correspondre à une variation de ce descripteur, puisqu'elle viole l'indication fournie pour le nombre. Les indices entre chevrons indiquent un partage de valeur, ici du genre et du nombre entre la tête et ses modifieurs adjectivaux, ainsi qu'avec les constituants qui les dominent.

---

[89] L'état fonctionnel du ventricule gauche est crucial en cardiologie. Le ventricule droit ne revêt pas la même importance. *Fonction ventriculaire droite* n'est d'ailleurs pas un mot-clé du domaine.

[90] Dans cet arbre, nous avons laissé comme étiquette du noeud la catégorie du constituant. Nous aurions aussi pu la représenter comme un trait aditionnel : {catégorie=SN...}.

15.3.2 Engendrer des variantes possibles de termes

Une des variations possibles d'un terme de structure [SN Nom [Sadj [Sadj Adj] Adj]] est la modification du syntagme adjectival par un nouvel adjectif à gauche ou à droite. Pour le terme choisi, cela signifie qu'il est *a priori* possible d'en rencontrer la modification suivante :

[SN [Nom fonction][SAdj [Adj x][Sadj [Sadj [Adj ventriculaire]] [Adj gauche]]]

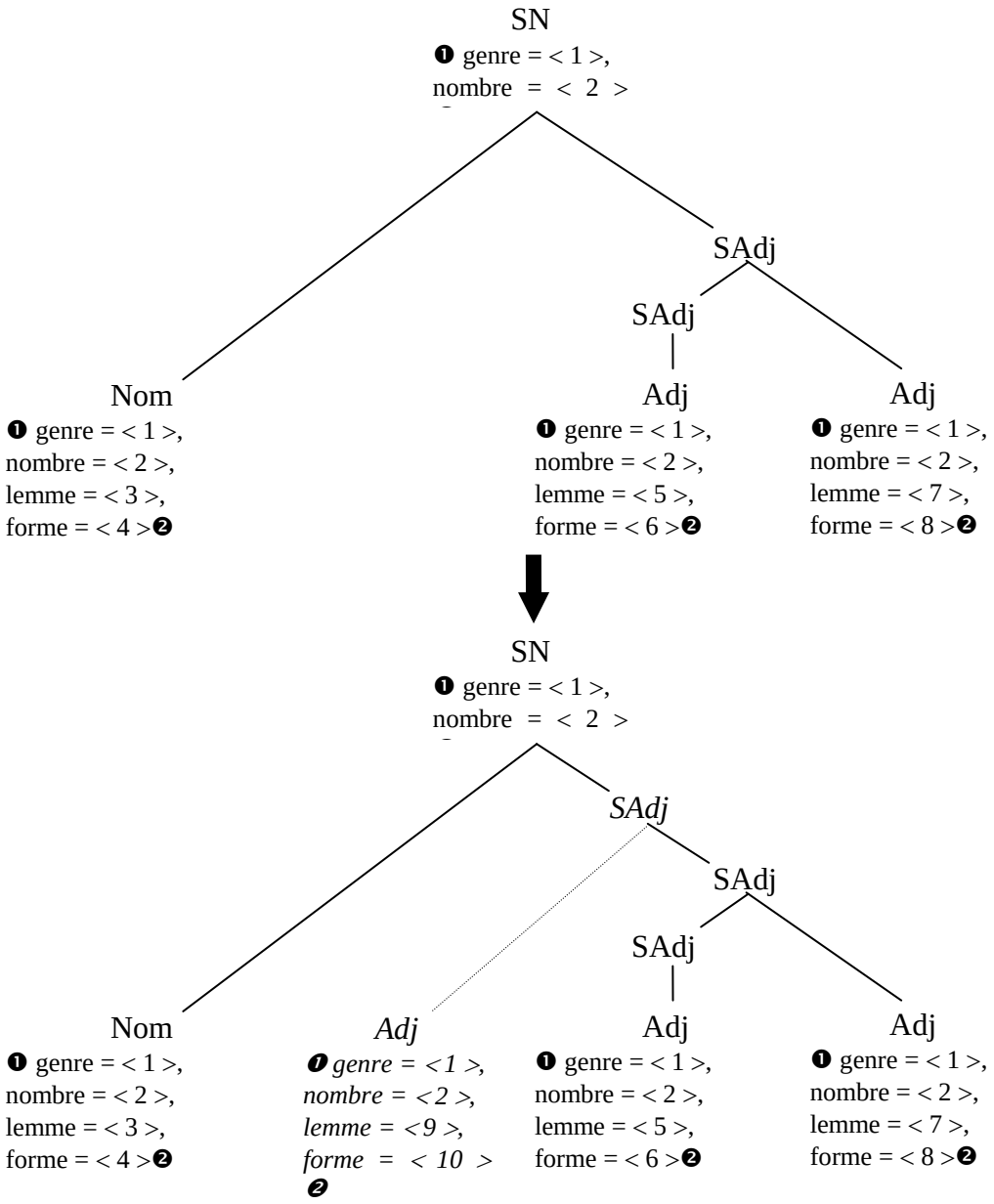
ou bien encore :

[SN [Nom fonction][SAdj [Sadj [Sadj [Adj ventriculaire]] [Adj gauche]][Adj x]]]

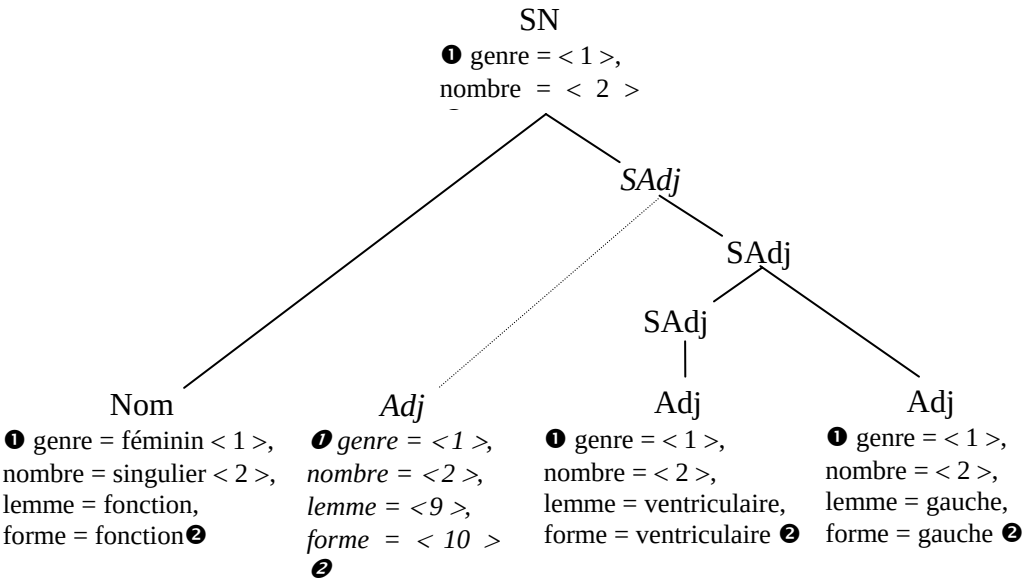
où *x* peut être remplacé par un adjectif quelconque. Les séquences correspondantes sont *fonction* **x** *ventriculaire gauche et fonction ventriculaire gauche* ***x***, dans lesquelles **x** doit être un adjectif.

Des méta-règles servent alors à stipuler les transformations que peuvent éventuellement connaître les descripteurs. Elles prennent en entrée un arbre décrivant un descripteur et produisent en sortie un autre arbre représentant une variation possible de ce descripteur.

La méta-règle suivante :



appliquée à l'arbre représentant le descripteur *fonction ventriculaire gauche* produit l'arbre suivant :



Cet arbre correspond à l'interposition possible d'un adjectif entre *fonction* et *ventriculaire gauche*. Cet adjectif doit s'accorder avec *fonction*. C'est le rôle des indices entre chevrons sur les traits attachés aux noeuds : le trait nombre et le trait genre de l'adjectif inséré doivent avoir la même valeur que les traits correspondants attachés à *fonction*. Le lemme de l'adjectif ajouté n'est pas précisé par contre.

Les méta-règles comprennent donc des « décorations » sur les nœuds. Ces informations permettent de contraindre leur application. On pourrait ajouter par exemple le trait {type = relationnel ∨ qualificatif/relationnel} pour empêcher l'engendrement d'une variation avec un adjectif qualificatif : **fonction satisfaisante ventriculaire gauche*. L'adjectif *satisfaisante* portant le trait {type=qualificatif}, il y aurait conflit entre la valeur du trait dans la méta-règle et celle de *satisfaisante*.

Une autre méta-règle peut faire fond sur la valeur du trait nom-base, associé à *ventriculaire* pour engendrer l'arbre correspondant à *fonction du ventricule gauche*, où l'adjectif relationnel *ventriculaire* est remplacé par le syntagme prépositionnel équivalent. Cette transformation peut opérer dans l'autre sens, ce qui permet d'obtenir *infarctus myocardique* à partir d'*infarctus du myocarde*. Ces transformations sont donc conditionnées par la présence de certains traits. Le terme *infarctus du myocarde* peut être transformé en *infarctus myocardique* parce qu'est associé au nœud correspondant à *myocarde* le trait {adjectif-relationnel=myocardique}. Le terme *angine de poitrine* ne pourra pas être transformé de la même manière : l'adjectif *poitrinaire* a le sens d'« atteint de tuberculose poitrinaire » et n'est pas l'adjectif relationnel qui serait nécessaire pour le déclenchement de cette méta-règle[91].

Une méta-règle peut, dans des conditions bien définies, s'appliquer sur les résultats d'autres méta-règles. Les deux méta-règles vues précédemment peuvent par exemple se combiner pour engendrer la variation potentielle *fonction* {adjectif} *du ventricule gauche*.

[91] Pour ajouter ces contraintes, on associe à poitrine le trait {adjectif-relationnel=sans} et à poitrinaire le trait {nom-base=sans}, par exemple.

C. Jacquemin a mis au point par expérimentation sur différents corpus les méta-règles nécessaires pour rendre compte des transformations effectivement rencontrées pour les termes techniques de plusieurs corpus techniques (médecine, métallurgie ...). Toutes les variantes potentielles prévues par les méta-règles et leurs combinaisons à partir d'un ensemble de descripteurs du domaine sont engendrées.

15.3.3 Repérage des variations syntaxiques engendrées

L'analyseur robuste FASTER, développé par C. Jacquemin, recherche ces variations dans un corpus du domaine le plus souvent étiqueté au préalable. C'est un analyseur très particulier : il se cantonne à un type de composant syntaxique, le groupe nominal, et s'en tient aux groupes qui comprennent certaines entrées lexicales, dans des relations de dépendance bien définies et obéissant à des contraintes fines grâce aux traits décorant les nœuds non-terminaux. Dans ***Menelas***, les méta-règles appliquées à *fonction ventriculaire gauche* permettraient de repérer *fonction systolique ventriculaire gauche, fonction ventriculaire gauche systolique* [92], ainsi que *(évalution de la) fonction globale du ventricule gauche* et *fonction du ventricule gauche*. Les transformations non « prévues » aboutiraient à un silence, c'est-à-dire à la non-extraction d'une variation effective. C'est le cas de l'acronyme, attesté : *FVG*. C'est le cas encore du remplacement de la tête par un hyponyme *: cinétique ventriculaire gauche* ou par une périphrase : *état fonctionnel du ventricule gauche*.

15.3.4 Vers une grammaire de la variation terminologique

C. Jacquemin distingue au sein des variations possibles les modifications (la tête ou un dépendant reçoit un modifieur : *fonction systolique ventriculaire gauche*), les permutations (*fonction ventriculaire gauche* / *fonction du ventricule gauche*) et les coordinations (comme l'hypothétique °*fonction ventriculaire gauche et droite*). Le tri des variations rapportées par l'analyseur entre variantes effectives et « bruit », séquences non reliées au terme de départ, manifeste une dissymétrie de ces trois opérations. La coordination, avec ses contraintes sémantiques, débouche souvent sur des variantes non ambiguës. La modification isole des séquences au statut plus incertain. La permutation enfin aboutit à un taux de bruit encore plus important : il tient au rôle sémantique flou des prépositions dites incolores, en français comme en anglais (*de*, *à*, *of*).

Ce sont là les premiers éléments d'une véritable grammaire de la variation terminologique, capable de caractériser précisément les opérations possibles et leur domaine d'application. On peut même se demander si, à côté de mécanismes très généraux intervenant dans les

[92] Phénomène d'incertitude positionnelle assez fréquent dans ce domaine. En voici un autre exemple : *syndrome douloureux thoracique* / *syndrome thoracique douloureux*.

différents langages spécialisés, ne peuvent pas se rencontrer des régularités particulières à tel ou tel domaine. Dans l'immédiat, cependant, il y a peu de différences d'un corpus à l'autre sur les types de méta-règles à utiliser, ce qui pourrait plaider pour une certaine stabilité de la langue technique au regard des mécanismes syntaxiques employés.

15.4 La recherche de candidats termes

Les deux approches que nous venons de présenter cherchent les variations d'expressions toutes faites de la langue générale ou de termes de langues de spécialité. On part donc de séquences répertoriées dont on cherche en corpus des réalisations modifiées. Le travail que nous examinons maintenant est orienté par l'objectif complémentaire, l'acquisition terminologique, c'est-à-dire repérer les termes d'un domaine quelconque qui n'ont pas encore été répertoriés. Il s'insère dans un contexte industriel, la Direction des Etudes et Recherches d'Electricité de France (DER-EDF).

Une grande entreprise industrielle comme EDF doit maîtriser des flux d'informations électroniques immenses : rapports de recherche internes, articles et publications glanées sur les réseaux, documents destinés au public, etc. Il importe de pouvoir rapidement retrouver l'information pertinente dans cette masse de données, par exemple extraire les documents qui parlent d'une notion donnée.

Pour certains domaines, une terminologie a été établie par des documentalistes ou des terminologues. Elle répertorie les principales notions du domaine et leurs réalisations linguistiques : les termes correspondants. Elle comprend éventuellement des liens de synonymie, d'antonymie, d'hyperonymie. Par exemple, on trouvera dans la terminologie du domaine du TALN des termes comme *analyseur syntaxique*, *formalismes d'unification*, *chaînes de Markov*, un lien de synonymie entre *analyseur syntaxique* et *parseur*, un lien d'hyperonymie entre *parseur* et *analyseur robuste* (un *analyseur robuste* est un type de *parseur*). Ces liens sont utilisés pour élargir les recherches effectuées : un système de recherche d'information pourra, grâce à cette terminologie, rapatrier les textes parlant de *parseur* et d'*analyseur robuste* si la demande porte sur les *analyseurs syntaxiques*.

Dans d'autres domaines, il n'y a pas de terminologie disponible. Cette absence peut tenir au coût de la constitution d'une terminologie par des documentalistes. L'évolution extrêmement rapide de certains secteurs peut aussi contrecarrer le dessein de prendre un « instantané » des termes qui y sont employés : l'image produite a toutes chances d'être déformée. Le vocabulaire de la navigation sur les réseaux (Internet, Web) offre un bon exemple de tels changements incessants. L'acquisition terminologique a de manière générale pour objectif d'isoler les dénominations d'un domaine, pour créer ou compléter une terminologie.

D. Bourigault a développé à la DER-EDF Lexter (Bourigault, 1993), un analyseur destiné à isoler les « candidats-termes » présents dans un

corpus de texte « tout-venant », préalablement étiqueté. Il entend par candidats-termes les syntagmes nominaux qui ont un fonctionnement dénominatif. L'hypothèse fondamentale est qu'un analyseur peut « dégrossir » le travail de repérage des dénominations effectives d'un domaine. Clairement, certaines séquences nominales, parce qu'elles font référence au cotexte ou au contexte, n'ont pas la généricité requise pour des dénominations (Kleiber, 1984). Par exemple *le maintien de sa température* ne serait pas retenu, en raison du possessif, tandis que *le maintien de température*, voire *le maintien de la température* le seraient : le déterminant zéro et le déterminant défini sont compatibles avec une lecture dénominative.

15.4.1 Isoler les groupes d'allure dénominative

La première étape du travail de Lexter consiste à isoler les groupes nominaux d'allure dénominative « maximaux ». L'approche retenue ne s'appuie pas au premier chef sur des règles de structuration du groupe nominal en français. Il s'agit au contraire au départ de repérer les frontières, c'est-à-dire les catégories et suites de catégories qui forment les bornes, exclues, d'un tel constituant. Dans la séquence (*ibid.* p. 108) :

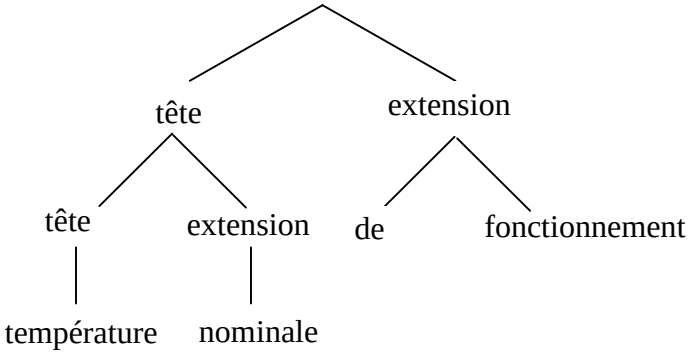
> le circuit d'aspersion de l'enceinte de confinement assure le maintien de sa température nominale de fonctionnement après une augmentation de pression

les éléments *assure*, *de sa*, et *après une* sont considérés comme des frontières. Le verbe est la limite d'un groupe nominal ordinaire. Par contre, *de sa* ne peut servir à articuler deux parties d'une dénomination complexe,

> Tête : température nominale
>
> Tête : température
>
> Expansion : nominale

après une non plus. On voit donc se superposer deux types de contraintes : l'une qui cherche à isoler les groupes nominaux, l'autre qui au sein de ce type de constituant, filtre ceux qui peuvent constituer des dénominations. Les groupes retenus sont *: circuit d'aspersion de l'enceinte de confinement*, *maintien, température nominale de fonctionnement*, *augmentation de pression*.

La deuxième étape ne garde que les groupes complexes : *maintien* est laissé de côté à ce stade. Les groupes sont en effet moins ambigus et apportent davantage d'information. Que l'on compare *données* et *base de*



données ou *analyse de données*. La première expression renvoie à l'informatique, la seconde aux statistiques, *données* tout seul potentiellement aux deux. À cette étape, les groupes sont également décomposés de manière récursive selon un schéma dépendanciel en Tête / Expansion[93]. La représentation de *température nominale de fonctionnement* est alors :

L'intérêt de ce type de décomposition, c'est de permettre les regroupements paradigmatiques qui sont si révélateurs en langage spécialisé. Regroupement sur les têtes : on peut mettre à jour des liens de co-hyponymie (entre plusieurs candidats-termes commençant tous par *analyseur* : *analyseur morphologique*, *analyseur syntaxique*, *analyseur robuste*, *analyseur montant* ...) ou d'hyperonymie (entre une séquence courte : *analyseur syntaxique* et une séquence qui la prolonge : *analyseur syntaxique déterministe*). Regroupement sur les expansions : il permet de voir les attributs à « spectre étroit » (qui modifient un nombre restreint de têtes : *déterministe* ne modifie guère qu'*analyseur* en TALN) et ceux qui sont moins spécifiques (*automatique* en informatique ou en TALN).

15.4.2 Le corpus comme norme

Les deux étapes reposent sur un postulat sous-jacent : limiter autant que possible l'appel à un savoir linguistique sur la langue dans son ensemble. Lexter nécessite seulement que le texte analysé ait été étiqueté préalablement pour pouvoir procéder à une analyse syntaxique partielle. Mais Lexter n'utilise ni informations sémantiques ni données de sous-catégorisation : prépositions régies par des noms prédicatifs ou par des adjectifs attendant un régime prépositionnel (*oublieux de*, *attentif à*, etc.). Cet « ascétisme » volontaire s'explique par la conviction, étayée par l'analyse détaillée de textes de domaines techniques distincts, qu'on ne peut pas forcément projeter les connaissances linguistiques générales sur les textes techniques, ou qu'inversement, les textes d'un domaine donné peuvent posséder des particularités combinatoires (des régimes de noms ou d'adjectifs qui le caractérisent) distinctes de celles d'un autre domaine.

Le corrélat logique de ce minimalisme est l'appel à l'apprentissage endogène. C'est considérer le corpus comme sa propre norme, et utiliser les régularités qu'il manifeste pour effectuer découpages et structuration. Lexter est souvent confronté à des ambiguïtés structurelles. Dans la séquence de ***Menelas*** *angine de poitrine instable*, faut-il rattacher *instable* à *poitrine* ou à *angine* ? Un locuteur français doit faire un effort pour simplement percevoir la difficulté. Un étranger qui ne connaîtrait que les mots isolés et pas le terme médical *angine de poitrine* partagerait pourtant l'hésitation de l'analyseur. Lexter dans un premier temps propose les deux découpages pour cette séquence : *[angine de poitrine] instable* et *angine de [poitrine instable]*. Le programme effectue un seul découpage pour les séquences non ambiguës. Dans un deuxième temps, Lexter regarde si l'un des sous-groupes des séquences ambiguës constitue un groupe non ambigu relevé au cours du premier temps. C'est ainsi qu'on rencontre

[93] Ce terme générique recouvre, comme *dépendant*, *modifieur* et *argument*.

dans ***Menelas*** *angine de poitrine*, mais pas *poitrine instable*. On choisit alors le découpage qui contient le groupe non ambigu, ici *[angine de poitrine] instable*. L'évaluation empirique de cette méthode sur différents corpus (*ibid.*, p. 113-114) donne les résultats suivants : dans 75 % des cas, la désambiguïsation obtenue est correcte ; 20 % des séquences restent non désambiguïsées ; 5 % des séquences sont désambiguïsées de manière erronée. Une comparaison de cette approche par apprentissage et d'une résolution des ambiguïtés par des règles *a priori* (Habert *et al.*, 1997) semble donner l'avantage à la première méthode. La délimitation des groupes maximaux repose également partiellement sur l'apprentissage. Certaines séquences constituent en effet des « frontières élastiques », c'est-à-dire qu'elles peuvent tantôt délimiter des groupes nominaux dénominatifs tantôt en faire partie. C'est le cas de *sur* + {article défini} (*ibid.*, p. 109-111). En général, c'est une limite :

1. on raccorde le câble d'alimentation sur le coffret de décharge batterie

Mais ce n'est pas toujours le cas :

2. action sur le bouton poussoir de réarmement
3. action sur le système d'alimentation de secours

En faire une limite intangible, c'est éliminer 2 et 3. L'accepter au sein des candidats-termes conduit à isoler *le câble d'alimentation sur le coffret de décharge batterie*, qui ne constitue certainement pas une séquence dénominative. La solution réside là encore dans l'apprentissage endogène. Il porte cette fois-ci sur les noms suivis d'une séquence sur + {article défini} + contexte droit immédiat. Un premier passage sur le texte relève tous ces contextes. Un second les trie et répartit les noms en deux groupes : ceux qui sont « productifs » avec *sur* (qui figurent dans le texte avec un nombre suffisant d'expansions différentes introduites par *sur* + {article défini}) et ceux qui ne sont suivis qu'exceptionnellement par *sur* + {article défini}. Lexter considère que l'expansion des premiers peut être introduite par *sur* + {article défini} et garde alors les séquences ayant pour tête à un niveau quelconque ces noms suivis d'une expansion introduite par *sur* + {article défini}. Dans les autres cas, *sur* + {article défini} continue à constituer une frontière. L'apprentissage porte donc ici sur des formes de sous-catégorisation.

15.4.3 Vers une grammaire des dénominations complexes possibles

L'ensemble retenu par Lexter est encore nettement trop vaste par rapport à ce qu'un expert du domaine considèrerait comme termes effectifs. Toutefois, l'objectif visé n'est certainement pas une automatisation totale de la mise en évidence des termes d'un domaine. Pour deux raisons fondamentales. La première, c'est que l'utilisation de Lexter sur des corpus variés, de domaines distincts, montre que les règles de bonne formation de termes possibles ne sont pas forcément les mêmes d'un domaine à l'autre. C'est pourquoi l'apprentissage endogène est justement

incontournable. La seconde raison tient à la complexité des mécanismes par lesquels une communauté langagière sélectionne, parmi les dénominations possibles, celles qui deviennent des dénominations effectives. Si l'on peut espérer diminuer la taille de l'ensemble des candidats-termes extraits d'un corpus, repérer ceux d'entre eux qui fonctionnent réellement comme des termes semble difficilement automatisable.

Lexter matérialise, par les séquences qu'il considère comme des bornes, un certain nombre d'hypothèses sur ce qui ne peut pas figurer dans une séquence nominale pour qu'elle puisse être employée comme une dénomination. La démarche suivie dans la mise au point et le test du logiciel sur des corpus variés ont conduit à rajouter d'autres règles, également négatives. La démarche est proche de celle utilisé pour l'étiquetage (cf. chapitre VIII) : peu à peu, on dégage les régularités à l'œuvre et on met au point des procédures qui s'appuient sur elles. Au total, Lexter, au delà des procédures mises en œuvre, essaie donc de formaliser partiellement la notion de dénomination possible.

15.5 Enjeux pratiques et théoriques

15.5.1 Améliorer la description lexicographique

Barkema (*ibid.*) souligne que le degré de flexibilité d'une expression est rarement indiqué par les dictionnaires qui donnent cette séquence. Le dictionnaire *COBUILD* (Sinclair *et al.*, 1987) fait partiellement exception : pour *moment of truth* (*minute de vérité*), est ainsi indiqué que la seule modification possible de l'expression est l'utilisation au pluriel. Le repérage des réalisations possibles tel qu'il est effectué par Barkema permet d'enrichir la description lexicographique des expressions concernées. Il en va de même en terminologie spécialisée, où les résultats de FASTER isolent des variantes à intégrer dans les ressources lexicales.

En acquisition terminologique, Lexter permet d'enrichir le répertoire des termes à utiliser. L'écrémage par ce programme des dénominations possibles facilite le travail du lexicographe spécialisé. Les concordances de mots fréquents sont en effet souvent très difficiles à dépouiller et à organiser. Le découpage opéré en tête / expansion et les regroupements par têtes et par expansions offrent au contraire une vision synthétique du fonctionnement syntagmatique et paradigmatique des noms « pivots » du texte étudié. L'un des résultats de Lexter est d'ailleurs un réseau terminologique hypertextuel. Chaque candidat-terme est relié à sa tête et à son expansion et d'autre part à tous les candidats-termes dont il est lui-même tête ou expansion. Le lien aux documents de départ permet de replonger les séquences extraites dans leur contexte. Le tout permet à un connaisseur du domaine de séparer dans de bonnes conditions les termes effectifs des groupes parasites.

L'acquisition terminologique, possible avec FASTER, réalisée avec LEXTER, est une tâche dont les résultats sont difficiles à évaluer

objectivement. Il n'existe pas de corpus de tests où les termes pertinents seraient isolés et qui serviraient ainsi d'aune pour mesurer l'apport de ces outils. En outre, le projet de créer de tels corpus est peut-être chimérique. Ce sont des ensembles de termes distincts qui risquent d'être repérés par des experts différents en fonction de leurs préoccupations et de leurs points de vue. Un spécialiste d'épidémiologie et un cardiologue n'identifieront pas forcément les mêmes séquences dans ***Menelas***.

15.5.2 Distinguer variantes et variations

Dans les recherches de Barkema comme dans celles de C. Jacquemin, une fois repérées des variations autour de séquences de départ, termes ou expressions toutes faites, une des difficultés consiste à isoler les variantes effectives, celles qui fonctionnent comme des réalisations possibles pour les expressions considérées.

En langue de spécialité, c'est le recours à un expert qui seul permet de trancher. En langue générale, il faut éliminer les variations qui constituent des défigements intentionnels, des jeux de langage[94], et non des variantes des expressions de départ. Ainsi, les exemples 7 (*a not-so-cold war against Kaddafi*) et *9 (a period of cold and hot civil war which ended with Hitler's invasion of Austria*) de Barkema semblent relativement éloignés du sens originel, qui renvoie au monde d'après Yalta et qu'évoquent les exemples 5, 6 et 11 par exemple.

15.5.3 Importance quantitative de la variation

C. Jacquemin a évalué les résultats de l'extraction de variations de descripteurs engendrées par méta-règles. Il a utilisé à l'INIST[95] un corpus de 125 000 mots dans le domaine de la physique de la métallurgie et un sous-ensemble du lexique terminologique PASCAL utilisé à l'INIST pour l'indexation manuelle (6 621 termes liés à la physique et à la chimie de la métallurgie). Les méta-règles étaient au nombre de 112. Les occurrences de termes et de leurs variantes couvrent (en nombre de mots) 7 % de la surface du corpus, les variantes représentant 28 % de cette « zone terminologique ». Les variantes validées représentent 15 % des occurrences de termes. Cette estimation semble d'ailleurs une valeur plancher, au regard d'expériences sur d'autres langues et d'autres corpus. La variation terminologique est donc loin d'être négligeable, contrairement à un préjugé répandu : les termes seraient les « noms » univoques et stables des notions d'un domaine. Les résultats de Barkema vont dans le même sens, cette fois-ci pour la langue générale. Il semble en effet qu'au total, l'intuition linguistique ou, en langage spécialisé, celle d'un terminologue voire d'un expert du domaine, sous-estime les variantes effectives des dénominations complexes. Le recours au corpus renouvelle

[94] Cf. (Authier-Revuz, 1995).
[95] Institut National pour l'Information Scientifique et Technique - CNRS.

donc l'analyse de la variation de ces unités polylexicales.

15.5.4 Caractériser la flexibilité « normale »

Barkema distingue (*ibid.*, p. 40-41 ; 1993) trois dimensions qui s'articulent : la flexibilité syntaxique : la possibilité pour un groupe de se voir appliquer tout ou partie des règles du constituant dont il relève, la compositionnalité : le fait que le sens de la séquence soit ou non fonction du sens de ses constituants, et enfin la « collocativité » : les préférences d'emploi d'un mot (comme dans l'association privilégiée *économiste* et *distingué* où l'adjectif peut être modifié, coordonné, etc.). Les travaux sur le « figement » ont sans doute eu tendance à confondre ces dimensions qui sont partiellement indépendantes. Le recours aux corpus permet de cerner précisément la première d'entre elles.

Barkema montre comment un corpus arboré permet de fournir une caractérisation fine de la flexibilité attendue pour un schéma syntaxique donné. On peut alors porter un jugement sur les réalisations effectives d'une expression relevant de ce schéma. L'emploi d'un corpus arboré souligne le fait que certaines réalisations d'un schéma syntaxique sont plus probables que d'autres, pondérations qui échappent pour l'essentiel à la conscience d'un locuteur. Les contraintes sur la flexibilité ont suscité depuis longtemps les recherches. Barkema essaie de caractériser précisément l'autre pôle de l'opposition : la flexibilité « normale ». C'est effectivement une tâche nécessaire pour pouvoir parler en connaissance de cause de degré de figement. Le corpus offre le moyen de pondérer les règles applicables à un constituant donné.

16. UTILISER DES PARSEURS ET DES CORPUS ARBORES

16.1 Utiliser des parseurs

La mise au point des parseurs nécessite des mécanismes complexes qui sont dans l'immédiat plutôt l'apanage d'informaticiens que de linguistes. L'écriture et l'ajustement de grammaires pour des analyseurs robustes nécessite par exemple des mécanismes de pistage (de trace, disent les informaticiens) : examiner en détail le processus même d'analyse d'une phrase, pour vérifier la pertinence des règles employées, ajouter des règles nécessaires, etc., comme dans le « banc d'essai » de grammaires de l'université de Nimègue (Nederhof et Koster, 1993, p. 174). Ou encore un générateur qui produit aléatoirement des phrases en fonction des règles et du lexique utilisés : cela permet de repérer certains incohérences ou le laxisme sur certains points de la grammaire.

L'utilisation de parseurs pour la constitution de corpus arborés suppose encore dans une coopération étroite entre linguistes et informaticiens. Les

exemples de telles coopérations sont encore rares : le groupe de Nimègue***, Lancaster Treebank*** (Black *et al.,* 1993) et ***Penn Treebank*** (Marcus *et al.,* 1993).

16.2 Utiliser des corpus arborés

Pour parler des corpus annotés syntaxiquement, on utilise également les dénominations de banques d'arbres (*treebank*) et de bases de données syntaxiques (*syntactic database*) (Souter et Atwell, 1994, p. 142). Ces appellations pourraient faire croire à une utilisation aisée des corpus arborés, au même titre que les bases de données du commerce. Il n'en est rien.

Au sens informatique, une base de données associe des tables d'information représentant des relations dans un sens assez proche de celui de la théorie des ensembles et des méthodes pour exprimer des requêtes sur les informations présentes dans une collection de tables, ces méthodes faisant appel à l'algèbre relationnelle qui permet d'exprimer ces requêtes sans entrer dans les détails de la mise en œuvre des opérations. Dans une « base de données syntaxiques *»,* il y a bien accumulation d'informations (et un certain « démembrement », puisque les analyses sont simplement juxtaposées). Mais n'y sont présents ni une formalisation générale des données présentes (on a déjà souligné l'éclatement des pratiques d'annotation syntaxique[96]) ni un langage de requête adéquat, ni même la possibilité d'ajouter ou de retirer des informations, ce que permettent les bases de données. La variété des informations présentes et leur structuration complexe (en termes d'enchâssement de constituants mais aussi de liens horizontaux — par exemple pour les co-références ou pour les discontinuités — ou encore de structures de traits décorant les nœuds) constituent, il est vrai, un défi à la formalisation.

C'est LDB (*Linguistic DataBase*) qui se rapproche le plus d'un outil de gestion et d'interrogation de vastes ensembles de phrases arborées. Cet outil a d'ailleurs été utilisé pour d'autres ensembles arborés que ceux de l'université de Nimègue pour lesquels il a été conçu[97]. Il est possible donc de transformer un corpus arboré pour le rendre interrogeable par LDB[98]. Halteren et Heuvel (1990) offrent une présentation approfondie de l'ensemble des manipulations offertes. Les interrogations peuvent associer les contraintes structurelles (un nœud de telle catégorie dans telle position de dominance ou de dépendance par rapport à tel autre nœud) et des conditions sur les « décorations » des nœuds qui peuvent comporter un certain nombre d'étiquettes (ce qui équivaut à un système de traits). On peut par exemple chercher les phrases à constructions bi-transitives (du type *I gave him a book*), ou encore construire un tableau indiquant le nombre de noms modifiés par un groupe adjectival préposé et leurs

[96] (Souter, 1993) le montre en détail sur 7 collections ou corpus arborés.
[97] Par exemple, pour la version parsée, au sein de l'équipe ASCOT de l'université d'Amsterdam, du *Longman Dictionary of Contemporary English* (*LDOCE*) (Souter, 1993, p. 204).
[98] Voire d'autres données arborescentes, comme des définitions de dictionnaire (Halteren et Heuvel, 1990, p. 10).

correspondants avec adjectif postposé, et le nombre de noms non modifiés. C'est LDB que Barkema a utilisé pour déterminer les différentes réalisations syntaxiques du patron de base adjectif nom singulier.

Comme pour l'étiquetage, deux grandes fonctionnalités sont nécessaires. Elles doivent d'ailleurs pouvoir se combiner. D'abord filtrer les arbres répondant à des contraintes arbitrairement complexes. Les outils actuellement disponibles (comme ceux fournis avec ***Penn Treebank***) sont encore rudimentaires et en tout état de cause non génériques : ils sont faits pour traiter d'arbres selon un format d'encodage donné et ne travaillent pas à un niveau de généralité suffisant. Deuxième fonctionnalité : transformer des arbres. Il peut s'agir de changer des étiquettes pour faciliter l'interprétation, ou de restructurer des sous-arbres. Alors que les techniques de transduction d'arbre sont bien maîtrisées en informatique, leur mise à la disposition des utilisateurs de corpus arborés reste pour l'essentiel à réaliser[99].

[99] Cf. (Habert *et al.*, 1997) pour une utilisation de la transduction d'arbres pour la comparaison de deux outils d'acquisition terminologique.

CHAPITRE III

LES RESSOURCES LEXICALES POUR L'ETIQUETAGE SEMANTIQUE

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

À l'heure actuelle, ces corpus porteurs d'annotations sémantiques n'existent cependant qu'à l'état embryonnaire[100]. Les expériences menées sont très diverses, reflets de conceptions sémantiques très différentes. L'essor des corpus arborés a fait suite à celui des corpus étiquetés et on peut s'attendre dans les prochaines années à l'apparition et au développement des corpus porteurs d'annotations sémantiques. Mais l'étiquetage sémantique est d'abord conditionné par la mise à disposition des connaissances sémantiques. La nature même des sources lexicales utilisées détermine en grande partie la méthode d'étiquetage et le jeu d'étiquettes retenus. Aujourd'hui, c'est donc la question de ces ressources qui paraît centrale.

Ce chapitre décrit les principales sources actuellement utilisées ou utilisables pour étiqueter sémantiquement des corpus. Seules les connaissances sémantiques sont prises en compte[101]. L'objectif est non pas de dresser un catalogue de ces ressources[102] mais d'en esquisser une typologie. Ces ressources ont été conçues selon des principes et dans des perspectives variées. Elles portent l'empreinte de ces différences de conception. Il s'agit ici d'évaluer dans quelle mesure elles peuvent servir à l'étiquetage sémantique de corpus et plus précisément à

[100] Ils ne dépassent guère 200 000 mots.
[101] Nous ne mentionnons donc pas les autres types de connaissances (phonétique, morpho-syntaxique…) que ces sources, les dictionnaires notamment, peuvent comporter.
[102] On trouvera ce type de catalogue sur des pages web régulièrement mises à jour. Un groupe de travail de l'Association for Computational Linguistics (ACL SIGLEX, Special Interest Group on the Lexicon) se charge notamment de recenser les ressources lexicales disponibles (http://www.clres.com/dict.html).

la désambiguïsation lexicale, même si ce n'est pas dans ce but qu'elles ont été conçues.

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

17. UN OBJECTIF: LA DESAMBIGUISATION LEXICALE

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

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

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

Dans la première version, à chaque mot est associée une étiquette reflétant le sens dans lequel il est employé : la distinction et la numérotation des sens est reprise du *Petit Robert*[103]. Dans ce cas, chaque mot est étiqueté[104]. Dans la deuxième version, en revanche, il s'agit d'un étiquetage partiel, qui ne concerne que les marques de modalités et qui devrait permettre d'observer la répartition de ces modalités dans l'ensemble du corpus. Comme au niveau syntaxique, ces étiquettes pourraient être complexes et combiner plusieurs traits.

Nous ne prenons ici en compte que le premier type d'étiquetage qui associe un ou plusieurs sens à un mot ou à une unité textuelle. On parle dans ce cas de *désambiguïsation lexicale*[105] (*word sense disambiguation*). Il faut entendre ce terme dans un sens technique. L'objectif est d'identifier le *sens* dans lequel un mot est employé. Concrètement, il s'agit en fait d'un numéro de sens, ce sens étant choisi dans une liste finie de sens,

[103] Dans l'édition de 1973. La valeur 0 indique que le mot a un sens unique.
[104] *Ne* ne porte pas d'étiquette sémantique parce qu'il n'a pas un fonctionnement autonome. Il forme avec *pas* un seul et même constituant discontinu.
[105] Lorsque le contexte est clair, nous parlons plus simplement de *désambiguïsation*.

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

18. UNE OPPOSITION FONDAMENTALE : CONSTRUCTION LEXICALE OU CONCEPTUELLE

Une première distinction oppose les bases lexicales aux bases conceptuelles : les premières décrivent des mots et les secondes des objets[106] du monde tels que nous nous les représentons.

Mettons cette opposition en évidence à partir d'un exemple. Le mot *fauteuil* et la notion ou le *concept*[107] de fauteuil sont deux choses différentes. Le concept se définit traditionnellement soit par l'ensemble des chaises du monde réel auxquelles il renvoie, soit plutôt par un ensemble des propriétés[108] : un fauteuil est ainsi un siège comportant généralement quatre pieds, un dossier et des accoudoirs, un siège étant lui-même un meuble fait pour s'asseoir. Si le mot *fauteuil* se définit en partie comme le concept auquel il renvoie, il se définit aussi en opposition à tout un ensemble de mots comme *siège*, *chaise*, *tabouret*, *bergère*, par les connotations de confort, d'aisance et d'importance qu'il véhicule (« arriver dans un fauteuil », « fauteuil de président »), par ses emplois métonymiques (le fauteuil de président désignant souvent la fonction de président), etc.

Dans la pratique, les bases lexicales et conceptuelles dessinent deux espaces différents. Leur structure est parfois similaire : la relation SORTE-DE (IS-A, en anglais) de l'Intelligence Artificielle et de ses réseaux sémantiques est le pendant conceptuel de la relation d'hyponymie entre les mots[109]. L'opposition est parfois difficile à caractériser : on voudrait distinguer des catégories conceptuelles universelles ou du moins indépendantes de la langue mais force est de constater qu'un francophone et un anglophone — sans parler des inuits ou des

[106] « Objet » est ici à entendre dans un sens large : il s'agit aussi bien d'objets concrets que d'entités abstraites ou d'événements.
[107] Nous ne parlons pas ici de notion mais de concept. Ce terme est utilisé en l'Intelligence Artificicelle pour désigner l'image mentale que nous nous faisons des entités du monde, sans préjuger de la nature de cette image ou de son rapport au monde « réel ».
[108] On oppose ainsi les définitions extensionnelles et intentionnelles.
[109] Cf. (Kleiber &Tamba, 1990).

mandchous — ne se représentent pas le monde de la même manière. Il reste que les bases lexicales et conceptuelles diffèrent dans leur visée : les unes décrivent le lexique ; les autres cherchent à modéliser le monde ou la représentation que nous nous en faisons. Les bases lexicales sont parfois utilisées pour construire des catégories sémantiques, et les bases conceptuelles pour décrire les mots, mais dans chaque cas ce n'est pas leur visée première.

18.1 Bases de connaissances lexicales

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

18.1.1 Dictionnaires

Les dictionnaires, qu'ils se présentent sous forme papier, sur support électronique ou qu'ils soient conçus pour le support électronique, qu'ils soient spécialisés ou de langue générale, contiennent les mêmes types d'informations sémantiques. La figure 3.1 ci-dessous en donne un exemple, tiré d'un dictionnaire électronique anglais[110].

Pour une langue donnée, les dictionnaires recensent les mots et les expressions considérées comme lexicalisées et donnent pour chacun une liste de sens, organisée en une arborescence de sens et de sous-sens. Chaque sens est décrit par une combinaison d'indications généralement optionnelles : une définition, un trait de domaine, des indications concernant le niveau de langue ou la modernité du mot, une liste de synonymes ou de renvois analogiques, des antonymes, des expressions ou tournures dans lesquelles entre le mot vedette, des phrases ou citations comme exemples d'emploi, ou même une ou plusieurs traductions possibles dans une autre langue[111]. La liste des sens pour un mot donné varie d'un dictionnaire à l'autre, leur description aussi. On a souvent souligné le nombre des définitions circulaires où deux ou plusieurs mots se définissent les uns par les autres, ainsi que le manque de cohérence dans la forme même des définitions ou l'ordre des indications. Il faut rappeler par ailleurs que les dictionnaires sont destinés à des locuteurs ayant déjà une bonne maîtrise de la langue dont ils ne fournissent qu'une description parcellaire. Il sont donc *a priori* peu adaptés aux traitements automatiques.

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

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

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

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

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

Figure 3.1.— Exemple d'entrée de dictionnaire : le nom credit[112]

18.1.2 Thesaurus

Les thesaurus constituent un deuxième type de base de connaissances lexicales[113]. Ils organisent la description des sens de mots de manière différente des dictionnaires de langue. Ces derniers proposent avant tout des définitions de mots alors que les thesaurus reposent sur une sémantique plus spécifiquement relationnelle et servent à « mettre une idée en mots » ou à « trouver le mot juste ».

Les thesaurus comporte généralement deux voies d'accès. Un accès par les mots : comme les dictionnaires, les thesaurus comportent des entrées. Mais aussi un accès par les idées ou notions : les thesaurus regroupent les sens de mots en grandes catégories sémantiques et s'apparentent en cela aux ressources conceptuelles. Les figures 3.2 et 3.3 illustrent ces deux aspects.

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

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

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

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

Figure 3.2.— Exemple d'entrée de thesaurus : le nom credit[114]

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

La structuration en catégories sémantiques est également exploitée pour l'annotation de corpus. Dans le ***Roget's Thesaurus***[115], plus de 30 000 mots sont réparties dans 1 000 catégories sémantiques (numérotées de #1 à #1 000), elles-mêmes organisées en cinq hiérarchies de faible profondeur (cinq niveaux au maximum) (cf. figure 3.3). On voit donc apparaître deux niveaux possibles de catégorisation : aux feuilles de la hiérarchie des regroupements lexicaux ; dans la structure, une catégorisation conceptuelle.

De fait, diverses expériences[116] ont montré l'intérêt que présentent les catégories sémantiques d'un thesaurus comme le ***Roget's*** pour la désambiguïsation lexicale.

[114] Cet exemple est emprunté au thesaurus de Merriam-Webster dans sa version en ligne : ***WWWebster Thesaurus***, 1997, http://www.m-w.com/thesaurus.htm (sept. 1997).
[115] Il s'agit du ***Roget's Thesaurus*** de 1911 dans sa version électronique, actuellement disponible à l'adresse http://ecco.bsee.swin.edu.au/text/roget/headings.html.
[116] Voir notamment (Grefenstette, 1996) ou (Yarowsky, 1992).

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

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

18.1.3 Terminologies

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

18.2 Bases de connaissances conceptuelles

Alors que les ressources lexicales structurent l'espace des mots, les *réseaux sémantiques* et *ontologies*, issus d'une autre tradition aussi ancienne que la lexicographie[117], reflètent une conceptualisation du monde. Il s'agit cette fois de recenser les « catégories d'objets » ou

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

concepts du domaine considéré et éventuellement de représenter leurs propriétés ainsi que les relations qu'ils entretiennent entre eux. Il en résulte des hiérarchies ou des réseaux de concepts.

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

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

18.3 Une opposition réelle mais floue

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

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

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

À l'inverse, en dépit de l'ambition parfois affichée, il paraît illusoire de croire à l'universalité de l'ontologie résultante et de penser qu'une

conceptualisation du monde puisse être indépendante de la langue de son concepteur. Concrètement, cette dépendance est en particulier marquée dans le fait que les nœuds et les relations d'un tel réseau conceptuel portent des étiquettes empruntées au langage naturel, ce qui conditionne et biaise l'interprétation.

19. UNE GRANDE DIVERSITE DE RESSOURCES LEXICALES

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

19.1 Des distinctions de sens plus ou moins fines

Les bases lexicales fournissent généralement des distinctions de sens fines. Le *Petit Robert*[118] liste douze sens pour le nom *cours*, répartis en 6 sens principaux. Le *Webster's Collegiate Dictionary*[119] distingue trois entrées pour le nom *bank* et au total seize sens différents. ***WordNet*** ou ***le Roget's thesaurus*** distinguent respectivement 8 et 20 acceptions pour le mot *credit.*

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

Les dictionnaires établissent des distinctions « homographiques »[120] (Guthrie *et al*., 1994), représentées soit par des entrées distinctes, soit par les premières divisions de sens. Ainsi, pour l'anglais *bank*, on peut différencier l'*établissement bancaire* et la *berge,* pour le français *cours*, on peut distinguer les sens de *écoulement* et de *enseignement,* sans pour autant prendre en compte toute la diversité des sens donnés par les dictionnaires. Les dictionnaires donnent par ailleurs des distinctions de domaine (médecine, législation, technique…) qui sont elles aussi exploitables dans la perspective de la désambiguïsation lexicale (Guthrie *et al*., 1991).

Ces distinctions grossières peuvent également être obtenues à partir de thesaurus. Il faut alors tirer parti du haut de la hiérarchie des sens. Ces bases lexicales sont généralement structurées comme un ensemble de hiérarchies distinctes, chacune étant dominée par une catégorie sémantique générale. Pour un mot, on peut ainsi distinguer des grandes familles de sens sur la base de l'appartenance des sens à l'une ou l'autre

[118] Dans l'édition de 1972.
[119] Dans la 9e édition.
[120] Il s'agit plutôt de grandes familles de sens que de « vrais » homographes, ces sens pouvant être dérivés les uns des autres.

de ces hiérarchies. C'est l'approche de R. Basili *et al*. (1997, p. 248) qui ne retiennent, pour travailler sur les verbes, que 15 grandes catégories de ***WordNet*** (perception, émotion, création, changement…) et ignorent les distinctions plus fines internes à chaque catégorie. Le verbe anglais *record* ou son équivalent français *enregistrer* admettent ainsi en langue générale, trois sens représentés par les catégories de la cognition, de la communication et de la perception. E. Agirre et G. Rigau (1996) exploitent de la même manière les 25 grandes catégories de noms de ***WordNet*** pour établir des grandes oppositions de sens. Dans (Bouaud *et al*., 1997), « une catégorisation à gros grain » est élaborée de la même manière à partir d'une nomenclature médicale dans la perspective d'un étiquetage sémantique de Menelas.

Si ces sources permettent de décrire des distinctions de sens fines ou grossières, il est généralement plus difficile d'établir des distinctions intermédiaires. Les distinctions et hiérarchies de sens des dictionnaires ou thesaurus ne reflètent pas une description homogène dans sa granularité. De fait, dans WordNet, certains liens hyponymiques reflètent une proximité sémantique beaucoup plus grande que d'autres : « [on trouve] des liens qui semblent représenter, pour certains, une courte distance (RABBIT-EARS IS-A TELEVISION-ANTENNA) et pour d'autres, une longue distance (PHYTOPLANKTON IS-A LIVING-THING) »[121] (Resnik, 1995a).

19.2 Des ressources générales ou spécialisées

Il faut également distinguer les sources qui permettent de décrire la langue générale et celles qui rendent compte d'une langue spécialisée[122].

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

Malheureusement, les sources spécialisées font souvent défaut et

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

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

celles qui existent ne peuvent pas être réutilisées dans une perspective différente de celle pour laquelle elles ont été conçues initialement. L'expérience de (Charlet *et al*., 1996) est instructive à cet égard. Travaillant dans le domaine médical où les expériences de ce type sont anciennes, ces auteurs ont cherché, pour modéliser le domaine des maladies coronariennes, à réutiliser une base de connaissances préexistante, ***Unified Medical Language System*** (***UMLS***, (Humphrey et Lindberg, 1989)), précisément conçue comme un réseau sémantique unifié pouvant être utilisé dans différentes perspectives. Cette tentative s'est soldée par un échec et les deux principales raisons invoquées ne sont en rien spécifiques à cette expérience. La première concerne la couverture du domaine. Même si UMLS est une base de connaissances spécialisée, les auteurs font un constat similaire à celui que fait R. Basili pour les ressources lexicales générales : ils ont du enrichir certaines parties de la hiérarchie. La seconde est plus fondamentale : l'ontologie d'un domaine dépend d'un point de vue sur ce domaine et de la tâche qui est visée et de la tâche pour laquelle elle a été conçue ; elle n'est donc réutilisable que dans la mesure où la tâche demeure la même, ce qui est rare[123].

Les ressources lexicales font donc particulièrement défaut lorsqu'on se propose de traiter des corpus spécialisés. Deux autres pistes sont explorées. La première consiste à spécialiser une source lexicale générale pour l'ajuster à un domaine de spécialité. R. Basili et ses collègues tentent ainsi d'adapter la taxonomie des verbes de ***WordNet*** à divers domaines spécialisés en se fondant sur l'information contextuelle apportée par un corpus représentatif du domaine considéré. Ils distinguent les sens de verbes selon leur appartenance aux 15 grandes catégories sémantiques de ***WordNet*** (changement, cognition, communication, contact, émotion…). Il s'agit de sélectionner, parmi les différents sens associés à un verbe donné, ceux qui sont pertinents dans le domaine et d'ajouter les sens spécialisés qui ne seraient pas représentés dans le réseau initial[124]. La seconde piste vise à constituer les ressources lexicales dont on a besoin. Cette construction peut être manuelle mais cela limite considérablement la finesse de la description. R. Basili *et al*. (1993a) décrivent une expérience de ce type : ils utilisent une quinzaine de catégories très générales (action, artefact, lieu, matière…) pour étiqueter des textes spécialisés. Elle peut également être automatique. Il s'agit alors d'acquérir des connaissances lexicales spécialisées à partir des corpus du domaine : de nombreux travaux se situent dans cette optique, nous y revenons au chapitre IV.

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

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

19.3 Des sources plus ou moins informatisées

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

19.3.1 Dictionnaires et thesaurus sur support électronique

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

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

Dès lors que le texte est sur support électronique, on peut facilement passer d'une entrée à l'autre ; par des algorithmes sur les chaînes de caractères, on peut trouver les mots ayant une terminaison commune, rechercher tous les mots dont les définitions contiennent un mot donné, etc. Cela permet de s'affranchir partiellement des limites des définitions mentionnées par G. Miller : on peut reconstituer une partie de l'information manquante dans l'entrée de *arbre* en recherchant les entrées qui comportent les mots *arbre* ou *plante* dans leurs définitions.

19.3.2 Ressources électroniques

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

électronique et les ressources électroniques en tant que telles, dont le codage est conçu pour faciliter l'accès par des traitements automatiques, pour expliciter le statut des informations données et donc en fournir les règles d'interprétation.

19.3.3 Ressources informatisées

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

S'affranchir du support papier, c'est d'abord s'affranchir de l'ordre linéaire. La structuration du dictionnaire en entrées distinctes, la numérotation des sens et les diverses marques typographiques étaient des premiers pas pour échapper à cette contrainte et donner un accès « direct » à certaines données. Pour autant, il n'était pas possible de consulter en parallèle plusieurs entrées d'un dictionnaire, de repérer des symétries, des parallélismes et plus généralement la structure sous-jacente à un ensemble de mots sans un long parcours de renvois en renvois et un patient travail de reconstitution. De la même manière, pour se faire une idée générale de la hiérarchie d'un thesaurus, il est important de pouvoir varier le niveau de description[125], une approche dynamique que ne permettait pas le support papier. L'outil informatique permet désormais de structurer les ressources lexicales sur d'autres bases et la multiplication des liens entre les différents éléments d'information autorise de nouveaux modes de consultation. ***WordNet*** en est un exemple intéressant (cf. section 4)[126].

En conséquence, les dictionnaires électroniques permettent de gagner en cohérence. Prenons pour seul exemple le travail effectué sur le français par I. Warnesson (1985) pour constituer, à partir de différentes sources traditionnelles, un nouveau dictionnaire des synonymes reposant sur une définition formelle de la synonymie comme relation d'équivalence[127]. La cohérence d'un tel dictionnaire en faciliter l'exploitation.

Dans ce domaine de la lexicographie, l'informatique a déjà induit de profonds bouleversements, avec notamment de nouveaux modes de navigation et de nouvelles possibilités d'exploration, mais il reste probablement à inventer de nouvelles formes de dictionnaires. On peut penser en particulier à des bases de connaissances intégrées et dynamiques, aux degrés de granularité et de spécialisation variables, qui puisse être reconfigurées en fonction des besoins et des parcours de l'utilisateur et offrir ainsi différents points de vue à l'utilisateur. Reprenons l'exemple de *credit*. C'est un mot polysémique, riche en connotations et son entrée dictionnairique est trop riche pour être facile à exploiter. Si

[125] Soit en faisant un « zoom » pour concentrer son attention sur une zone donnée soit au contraire en faisant abstraction d'un certain niveau de détail pour dégager une vue d'ensemble.
[126] « En termes de couverture, les objectifs de ***WordNet*** diffèrent peu de ceux d'un bon dictionnaire standard de langue. C'est dans l'organisation de cette information que ***WordNet*** prétend innover. » (Miller *et al.*, 1993, p. 1).
[127] Qui respecte les propriétés de symétrie, de transitivité et de réflexivité.

l'utilisateur s'intéresse au domaine économique et financier, la plupart des sens deviennent immédiatement caduques tandis que les détails du deuxième sens prennent de l'importance. On devrait ainsi pouvoir considérer une base de connaissances sous différents points de vue.

20. UN EXEMPLE DE RESEAU LEXICAL : WORDNET

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

20.1 Un projet ambitieux

Depuis 1985, un groupe de psycholinguistes et de linguistes de l'université de Princeton a développé une base de données lexicale selon des principes suggérés par des expériences et des recherches en psycholinguistique sur l'organisation de la mémoire humaine. Depuis cette date, ce projet a pris de l'ampleur ; il se poursuit encore de nos jours. Le réseau ***WordNet*** disponible aujourd'hui est la version 1.5. Il peut soit être consulté en ligne soit être importé[128].

20.1.1 Représenter les sens de mots

L'objectif de ***WordNet*** est de décrire comment les sens de mots ou concepts[129] — et non les mots eux-mêmes — s'organisent les uns par rapport aux autres. En ce sens, WordNet ressemble davantage à un thesaurus qu'à un dictionnaire. La théorie sous-jacente est une théorie différentielle[130] : un sens se définit par la place qu'il occupe dans le réseau, par les relations de proximité ou de contraste qu'il entretient avec les sens voisins. Partant de ce principe, un sens est représenté par un ensemble de synonymes : « Les ensembles de synonymes (*synsets*) n'expliquent pas ce que sont les concepts ; ils en posent l'existence. On suppose que les locuteurs anglais ont déjà acquis ces concepts et sont en mesure de les reconnaître à partir des mots listés dans le synset. » (Miller

[128] WordNet est disponible par ftp anonyme depuis ftp.cogsci.princeton.edu ou ftp.ims.uni-stuttgart.de (sept. 1997). Il existe en différentes versions pour Unix, PC Windows et Macintosh.
[129] La terminologie de ***WordNet*** identifie le sens d'un mot au concept sous-jacent.
[130] Ceci s'oppose aux approches constructivistes qui tendent à définir un sens en le décomposant en primitives de significations.

et al., 1993, p. 5-6). Considérons l'exemple du mot *credit* pour lequel huit sens sont identifiés dans ***WordNet***[131]. En voici trois:

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

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

Dans ***WordNet***, la synonymie est contextuelle : « deux expressions sont synonymes dans un contexte linguistique C si la substitution de l'une par l'autre dans C ne modifie pas la valeur de vérité. Par exemple, le fait de substituer *plank* à *board* modifie rarement la valeur de vérité dans des contextes liés à la charpenterie, mais cette substitution serait totalement inappropriée dans d'autres contextes de *board*[132]*. »* (*ibid*., p. 6).

20.1.2 Mettre les « sens » en réseau

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

WordNet est conçu comme un réseau lexical. Les synsets en sont les nœuds. Ils sont reliés entre eux par des relations d'hyponymie, d'antonymie, de méronymie[133], d'implication ou de dérivation morphologique[134]. La figure ci-dessous montre de manière simplifiée[135] comment le premier sens de *credit* (*crédit*) se situe par rapport aux synsets voisins : c'est un hyponyme de *asset* (*avoir*), un hyperonyme lointain[136] de *credit-card* (*carte de crédit*), un antonyme de *cash* (*argent comptant*)*.*

Les relations qui structurent ***WordNet*** n'ont pas toutes le même statut.

---

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

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

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

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

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

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

La synonymie joue un rôle central dans la mesure où elle est interne aux nœuds et constitutive des synsets. Elle s'oppose à toutes les autres relations, qui relient les mots les uns aux autres. Cela revient à distinguer deux niveaux de relations : les relations *lexicales*[137] qui relient respectivement entre eux les mots et les relations *sémantiques* qui relient entre eux les sens de mots, c'est-à-dire les synsets ou concepts.

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

20.1.3 Quelques chiffres

La taille du vocabulaire couvert suffit à donner la mesure de l'ambition qui a présidé à la construction de ce réseau. ***WordNet*** comporte[138] 95 600 unités lexicales différentes : 51 500 mots simples et 44 100 expressions (*collocations*). À ces mots sont associés quelques 70 100 sens différents. Le tableau 3.1 montre comment ces unités et sens se répartissent.

Tableau 3.1

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

[137] Nous reprenons ici la terminologie de ***WordNet***.
[138] Les chiffres que nous citons sont ceux que donnent (Miller *et al.*, 1993). Ce sont des approximations, ce qui explique l'inexactitude des totaux. ***WordNet*** continue de croître.

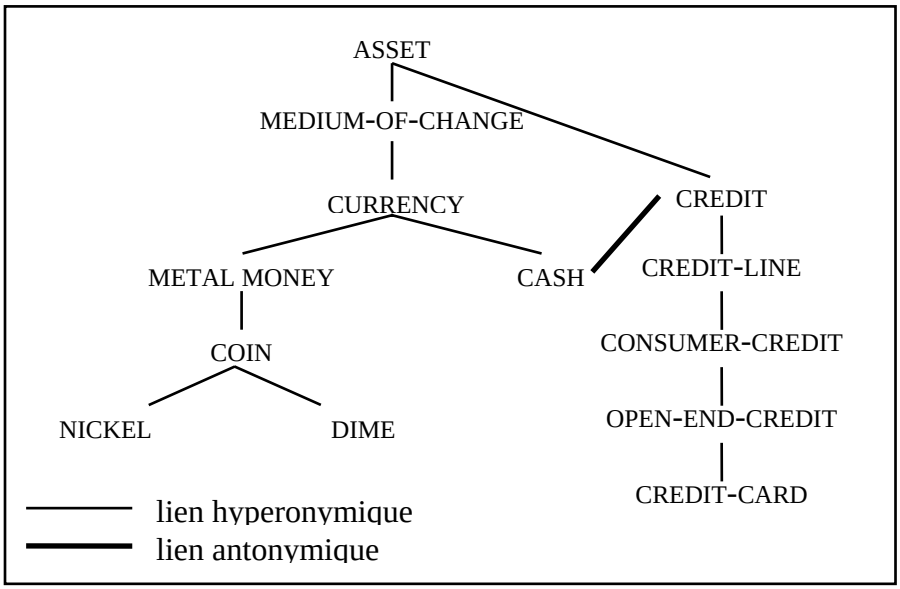


*Figure 3.4.— Exemple de sous-hiérarchie de **WordNet**.*

20.2 Une structure riche et différenciée

WordNet décompose le lexique en cinq catégories : noms, verbes, adjectives, adverbes et mots fonctionnels[139]. Chacune de ces catégories a sa propre structure interne. « Ce sont des expériences sur les associations de mots qui ont mis en évidence à l'origine que l'organisation […] varie d'une catégorie syntaxique à l'autre. » (*ibid*.).

20.2.1 Des hiérarchies de noms

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

televangelist < evangelist< preacher < clergyman < spiritual leader < person[140]

La structure induite est en fait un ensemble de 25 hiérarchies dominées

[139] Cette dernière catégorie n'est toutefois pas intégrée à ***WordNet*** (NDA).
[140] Dans x < y, le mot *x* est donné comme l'hyponyme du mot *y*. On aurait pour le français la séquence suivante : *télé-évangéliste* < *évangéliste* < *prédicateur* < *ecclésiastique* < *chef spirituel* < *personne*.

par des catégories sémantiques générales (*unique beginner*) : *person* dans l'exemple ci-desus ; *possession*, hyperonyme direct de *asset*, pour la sous-hiérarchie représentée par la figure 3.1. Au sein d'une hiérarchie, la hauteur est variable : selon les zones du lexique concernées, les synsets les plus bas se situent à 3, à 10, parfois même à 12 niveaux d'écart du sommet. De fait, si le vocabulaire technique se prête souvent bien à ce type d'organisation[141], il est plus difficile de définir des chaînes hyponymiques entre les mots de la langue courante (Kleiber et Tamba, 1990) : dans l'exemple ci-dessus, on peut se demander si tous les prédicateurs (*preacher*) sont effectivement des ecclésiastiques (*clergyman*).

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

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

Outre leur place dans cette structure hiérarchique, les sens des noms se définissent par des propriétés : leurs attributs, leur composition et leurs fonctions. La composition est décrite par différents types de relations méronymiques dans ***WordNet*** : les relations de composant à objet composé (*branche* / *arbre*), d'élément à ensemble (*arbre* / *forêt*) et de matière (*arbre* / *bois*). En revanche, les attributs (un arbre peut être grand, vieux…) et les fonctions (*une hache sert à couper…*) ne sont pas représentés dans ***WordNet***. Ce sont en effet des relations trans-catégorielles qui devraient à terme relier les hiérarchies de noms aux réseaux des adjectifs ou des verbes.

20.2.2 Des classes d'adjectifs

Les synsets d'adjectifs comprennent essentiellement des adjectifs qualificatifs[142], même si des noms ou locutions prépositionnelles utilisées comme modifieurs y figurent également. Ces adjectifs ne s'organisent pas comme les noms. Pour les adjectifs, il n'existe pas de relation hiérarchique

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

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

comme l'hyponymie.

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

Partant cependant du constat que certains adjectifs proches par le sens (*heavy* et *weighty*[143], par exemple) ont des antonymes différents (*light* et *weigthless*[144]) et que beaucoup d'adjectifs qualificatifs (*ponderous*[145]) n'ont pas d'antonymes directs, la structure retenue est celle de classes d'adjectifs similaires entre eux, ces classes étant organisées autour d'adjectifs pôles qui peuvent s'opposer à d'autres pôles par des liens d'antonymie. *heavy* et *light* sont donc considérés comme antonymes, mais *ponderous*, qui est similaire à *heavy* et qui n'a pas d'antonyme direct n'est qu'un antonyme indirect de *light*.

20.2.3 Des réseaux de verbes

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

La relation centrale pour le réseau des verbes n'est ni l'hyponymie, ni l'antonymie, mais l'implication. ***WordNet*** en distingue quatre types : la cause (*give* / *have* : *donner* / *avoir*), la présupposition (*succeed* / *try* : *réussir* / *essayer* ou *untie* / *tie* : *dénouer* / *nouer*), l'inclusion (*snore* / *sleep* : *ronfler* / *dormir* ou *buy* / *pay* : *acheter* / *payer*) et la troponymie[146] (*limp* / *walk*, *boiter* / *marcher*).

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

21. TABLER SUR L'EXISTANT

Les ressources lexicales existantes ont chacune leurs faiblesses. Dès lors qu'elles visent une couverture un peu large du lexique, elles reposent sur

[143] *lourd* et *pesant*, respectivement.
[144] *léger* et *de peu de poids*.
[145] *massif, pesant.*
[146] Un verbe x est un troponyme d'un verbe y si on peut dire que *x, c'est y d'une certaine manière.*

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

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

Ceci nous amène à souligner avec inquiétude l'absence de ressources similaires pour le français[147]. Si la recherche sur les corpus en français peut sans doute tirer profit de l'expérience anglo-saxonne pour éviter certains tâtonnements, des problèmes spécifiques se posent pour chaque langue, qui imposent certains ajustements, voire la mise au point de méthodes particulières ou le développement d'outils spécifiques. L'absence de ressources lexicales informatisée pour le français est déjà un frein pour tous les traitements sémantiques. Faute de moyens, la plupart des travaux français s'intéressent à l'acquisition de connaissances à partir de corpus (cf. chapitre VIII, section 5).

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

DEUXIEME PARTIE

DIMENSIONS TRANSVERSALES

CHAPITRE IV

DES MOTS AUX SENS : SEMANTIQUE EN CORPUS

22. DEFINITIONS ET ENJEUX

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

22.1 Un objectif commun : accéder au sens

Des corpus porteurs d'annotations sémantiques commencent à voir le

jour, mais on n'en est cependant qu'aux balbutiements, que ce soit pour la constitution de ces corpus ou pour leur exploitation.

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

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

22.2 Des applications variées

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

22.2.1 Analyse de contenu

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

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

Au delà de la seule étude du vocabulaire, l'ambition de M. Pêcheux avec l'analyse du discours est de mettre en évidence, sous la diversité des formes rhétoriques de surface, les phrases élémentaires ou « de

base » d'un discours. Il s'agit par exemple pour Pêcheux et ses collègues de mettre en évidence l'ambiguïté idéologique du « rapport Mansholt » (Maingueneau 1991).

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

22.2.2 Recherche documentaire

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

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

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

22.2.3 Acquisition de connaissances

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

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

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

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

23. CONSTRUIRE AUTOMATIQUEMENT DES ENTREES DE DICTIONNAIRE

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

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

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

23.1 Des ébauches d'entrées de dictionnaires

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

> <Nom vedette> :: [<données quantitatives>] <NOM DU CORPUS D'ORIGINE> ***Relat***.[148] <liste des noms voisins>. ***Vbs***. [149] <liste des verbes opérateurs>. ***Exp***. [150] <liste des expressions et de leurs expressions voisines>. ***Fam***. [151] <liste des variantes >.

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

> Growth :: [284 contexts, frequency rank : 25] MED ***Relat***. tumor ; effect, tissue ; antigen, protein, development. ***Vbs***. retard, stimulate, show, follow, enhance, accelerate. ***Exp***. growth

[148] Pour *related words.*
[149] Pour *verbs.*
[150] Pour *expressions.*
[151] Pour *family*.

hormone (cf. bone marrow, parathyroid hormone), growth rate (cf. growth retardation, folic acid), tumor growth (cf. body growth, tenuazonic acid), growth retardation (cf. dna content, body weight), body growth (cf. tumor growth, body weight).

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

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

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

Ces entrées ne ressemblent guère à des entrées habituelles de dictionnaire[152]. Pourtant, elles constituent un ensemble d'indications qui peut guider le lexicographe dans son travail de rédaction. Elles comportent six rubriques, les quatre dernières étant optionnelles.

23.1.1 Des données quantitatives

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

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

23.1.2 Le corpus d'origine

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

23.1.3 Les noms voisins

Cette liste, qui est introduite par le mot clef *Relat.*, comporte des noms donnés comme sémantiquement proches du nom vedette. Dans le corpus financier, *growth* se trouve au voisinage d'une dizaine de noms : *level, increase, gain **;** loss **;** performance, return, rise, decline, flow, expansion*. *S*oulignons la cohérence de cette liste[153]. Elle comporte essentiellement des synonymes ou des pseudo-synonymes (*increase, gain, rise, expansion*) et quelques antonymes (*loss, decline*). Même si le lien de *growth* avec *level*, *performance* et *flow* est moins évident, le rapprochement de ces termes paraît néanmoins assez judicieux. Seul *return* surprend. La liste des voisins est structurée en trois parties séparées par des points virgules. Sont ainsi distingués les voisins qui sont plus fréquents, aussi fréquents et moins fréquents que le mot vedette, cette indication pouvant refléter le degré de généralité. Pour le lexicographe, cette liste donne un premier aperçu des relations lexicales autour du nom vedette, relations dont il n'est pas évident de se faire une idée *a priori*, à la lecture du corpus ou même à partir de concordances. Cette liste doit être contrôlée, parfois émondée ou complétée : la liste des voisins de *year* semble peu satisfaisante, par exemple. Le retour aux contextes permet de vérifier le sens dans lequel les mots sont employés. Dans tous les cas, cette liste demande à être interprétée pour que soit identifiée la nature des relations lexicales sous-jacentes.

23.1.4 Les verbes opérateurs

Ces verbes sont introduits par le mot clef *Vbs*. Il s'agit des verbes auxquels le nom vedette est régulièrement associé, comme sujet, objet direct ou complément prépositionnel. Les verbes sont classés par ordre de fréquence décroissante. Cette rubrique renseigne sur les emplois du nom vedette et les relations dans lesquelles il entre. On constate ainsi que la croissance (*growth*) dans le corpus financier, est quelque chose dont le rythme évolue (*slow, accelerate, maintain, sustain, continue*), mais aussi quelque chose qui se prévoit (*expect*, *forecast*). En termes de fréquences,

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

c'est surtout quelque chose dont on parle ou qui donne des informations (*say*)[154]. En fait, cette rubrique des verbes opérateurs donne une première indication synthétique des contextes d'emplois du nom vedette. Le fait que *age* (*âgé de*) figure parmi les verbes associés à *year* explique la présence surprenante à première vue des noms de personnes (*women*, *child*, *patient*, etc.) aux côtés des termes de durée (*day*, *week*, *month*, etc.). C'est, semble-t-il, l'importance des contextes du type *woman aged of thirty years* qui rapproche *woman* et *year*.

23.1.5 Les expressions

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

23.1.6 Les variantes

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

Le recours aux corpus, plutôt qu'à l'introspection, est chose ancienne pour la lexicographie spécialisée et il est clair que les entrées ainsi constituées automatiquement demandent à être retravaillées par un

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

lexicographe. Le travail de G. Grefenstette montre cependant toutes les possibilités que le traitements automatiques de corpus ouvrent désormais. Rappelons en effet que les entrées données ci-dessus ont été engendrées de manière entièrement automatique. Ces entrées constituent des ébauches ou un premier dégrossissage qui donnent au lexicographe une vue synthétique sur le poids (données quantitatives) et le fonctionnement syntagmatique (expressions et verbes opérateurs) ou paradigmatique (voisins et variantes) d'un mot dans le corpus considéré.

23.2 Une méthode entièrement automatique

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

23.2.1 Une seule donnée, le corpus

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

La construction de ces entrées de dictionnaire ne fait appel à aucune connaissance sémantique. C'est là le point fort de la méthode qui repose sur des techniques de bas niveau (*knowledge-poor techniques*) en ce sens que le processus d'extraction repose entièrement sur des traitements morpho-syntaxiques et statistiques du corpus[155].

23.2.2 Un ensemble de traitements simples

Le traitement du corpus est effectué par le logiciel SEXTANT (Grefenstette, 1994a) qui traduit dans un premier temps le corpus préalablement

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

étiqueté en un ensemble de relations de dépendances syntaxiques. L'accent est mis sur les noms et ne sont conservées que les relations entre un nom d'une part et un adjectif, un verbe ou un autre nom, d'autre part. En simulant ce traitement sur les extraits de ***Menelas*** donnés ci-dessous, on obtient comme contextes pour le nom *épisode* ses relations avec les mots suivants[156] : *présenter* (OBJ), *survenir* (SUJ), *douloureux*, *précordial*, *hyperthermique*, *effort*, *repos*.

> Traité médicalement, il a déjà *présenté* à plusieurs reprises des ***épisodes*** *douloureux précordiaux d'effort* et de *repos*.
>
> Depuis cette époque on ne note aucune récidive d'angor jusqu'il y a 8 jours où il a *présenté* un ***épisode*** de *précordialgie survenant* à l'effort, durant environ 45 minutes, sans irradiation[157].
>
> On notait par ailleurs la *survenue* d'un ***épisode*** *hyperthermique*, probablement en rapport avec une mise en place prolongée d'une voie veineuse.

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

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

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

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

la comparaison des distributions tend à montrer que *épisode* est plus similaire de *symptomatologie* que de *crise*. Formellement, les contextes d'un nom constituent un ensemble de propriétés (ses attributs) et le

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

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

logiciel mesure le degré de similarité[158] entre deux noms sur la base du nombre d'attributs qu'ils partagent[159]. Dans la liste des voisins d'un nom vedette, on retient les noms qui en sont le plus similaires, à condition que, de manière réciproque, le nom vedette figure également en bonne position dans la liste des similaires de ceux-ci.

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

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

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

23.3 Les limites d'une approche empirique

Pour bien utiliser un outil comme SEXTANT dans une perspective lexicographique, il est également important d'en connaître les limites. L'approche décrite ci-dessus présente certaines faiblesses. La rubrique la moins satisfaisante est incontestablement celle des variantes qui mêle notamment les variantes orthographiques et dérivationnelles. L'algorithme de recherche des variantes morphologiques privilégie les variations qui ne portent pas sur l'initiale du mot et associe des mots qui ont seulement le

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

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

même préfixe (*antigen* est associé à *antibody* mais pas à *gene*)[160].

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

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

On touche là aux limites intrinsèques de l'approche présentée ici. Le travail lexicographique ne peut reposer entièrement sur les corpus. Mais si les informations extraites de corpus doivent être contrôlées, corrigées, complétées, elles constituent néanmoins une vue d'ensemble sur l'emploi d'un mot et une source importante pour la rédaction d'entrées de dictionnaire. Pour exploiter ce type de données, le lexicographe devra acquérir l'expérience des outils permettant de les obtenir, afin de dépister les points faibles de telle entrée, identifier les associations douteuses,

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

repérer les effets d'une analyse syntaxique inexacte ou ambiguë, et pour compléter les informations extraites par ses propres méthodes d'investigation.

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

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

24.1 Retrouver des textes dans une base documentaire

24.1.1 Principe général

Idéalement la requête de l'utilisateur spécifiant le type des documents recherchés devrait pouvoir être exprimée en langage naturel avec toute latitude dans le choix de la formulation ou, à la rigueur, dans un langage de requête, sous une forme explicite mais plus contrôlée. La formulation naturelle « *les textes décrivant les problèmes de circulation sur les grandes artères* » peut ainsi se traduire par une relation de localisation entre deux entités : LOCALISATION(problème de circulation, grandes artères). En pratique cependant, les systèmes commercialisés proposent généralement à l'utilisateur de formuler sa requête sous la forme d'une liste de mots clefs, éventuellement combinés par des opérateurs booléens (ex. circulation ET artères)[161].

Un système de recherche documentaire commence par indexer les documents de sa base, c'est-à-dire qu'il représente leur contenu sous la forme d'une liste de termes[162] représentatifs de ce contenu. Il extrait de la même manière des termes de la requête de l'utilisateur. Puis, il cherche à apparier les termes de la requête avec ceux d'un document pour évaluer la pertinence de ce document au regard de cette requête. L'objectif est bien entendu de retrouver tous les documents pertinents de la base et

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

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

ceux-là seulement. Dans la pratique, il faut trouver le meilleur compromis entre rappel et précision.

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

24.1.2 La question de la variation lexicale

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

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

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

Synonymie et polysémie sont en fait les deux faces du même problème : on voudrait fonder la recherche sur les sens de mots et non sur les mots eux-mêmes. Dans le domaine très actif de la recherche documentaire, c'est l'un des axes qui est exploré. Sans développer les problèmes liés à l'expansion de requêtes (Voorhes, 1994), les paragraphes qui suivent mettent l'accent sur la désambiguïsation lexicale

de gros volumes de textes tout-venant. S'il est trop tôt pour faire état d'expériences et de résultats sur des systèmes intégrant effectivement un traitement lexical, nous voudrions ici montrer l'une des pistes prometteuses, consistant à exploiter une base lexicale générale. Nous nous appuyons plus particulièrement sur le travail de M. Sussna (1993). Son impact sur un système de recherche d'information n'est pas réellement évalué mais il montre tout le parti qu'on peut tirer d'une base lexicale générale comme ***WordNet*** (voir chapitre III, *supra*).

24.2 Désambiguïser des corpus à l'aide de WordNet

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

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

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

L'approche de M. Sussna est par ailleurs contextuelle. Comme beaucoup de travaux de désambiguïsation lexicale[163], elle repose sur l'idée que le contexte d'un mot permet d'identifier le sens dans lequel il est employé. Sous-jacente est l'intuition que l'on tend à sélectionner pour un mot le sens qui est lié au contexte. De fait, dans la plupart des cas, nous ne percevons pas d'ambiguïté car le contexte suffit à réduire l'espace des sens possibles. L'idée est de retenir pour un mot donné le sens qui se rapproche le plus de ceux de ses voisins, c'est-à-dire de mesurer la parenté ou la distance sémantique[164] entre les sens de différents mots qui se trouvent contigus dans le texte et de retenir la combinaison qui

[163] Voir (Guthrie *et al*., 1994).

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

minimise la distance globale.

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

24.2.1 Un article désambiguïsé

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

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

C'est cette liste qu'il s'agit de désambiguïser en associant à chaque mot une étiquette spécifiant le sens dans lequel il est employé dans cet article. M. Sussna ne donne pas d'exemple de texte désambiguïsé mais nous proposons ci-dessous (point *f*) une version désambiguïsée de l'article *a*. Nous avons effectué cette désambiguïsation manuellement. Les étiquettes renvoient à des sens de ***WordNet*** (voir *supra* III.3). Le sens d'un mot est représenté par son numéro d'ordre dans la liste des sens possibles pour ce mot : c'est le 3^{e} des 6 sens de *strike* qui est employé ici. Ce sens est également décrit par le synset dans lequel il figure, *i.e*. l'ensemble de ses synonymes (entre accolades), ou à défaut, par la paraphrase (entre guillemets) donnée dans ***WordNet***[165].

***a*. texte original**

The allies after Nassau

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

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

***b*. liste de noms**

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

c*. liste de noms absents de *WordNet

Kennedy MacMillan skybolt

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

Nassau december U.S. NATO Europe Europeans Britain

***e*. liste de noms sélectionnés**

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

***f*. liste de sens**

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

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

Pour M. Sussna, l'objectif est donc de mesurer par une distance entre les nœuds de ***WordNet*** la proximité des sens de différents mots dans un espace sémantique, c'est-à-dire leur parenté[166].

[166] Cette question du calcul de la distance sémantique se pose dans les mêmes termes,

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

Pour simplifier, on peut donc, comme le font E. Agirre et G. Rigau (1996) ou P. Resnik (1995b)[167], ne considérer que la partie hiérarchique de ***WordNet*** : « Soit *C* l'ensemble des concepts dans une taxonomie organisée autour de la relation EST-UNE-SORTE-DE (IS-A) telle qu'un nœud puisse hériter de plusieurs pères. Intuitivement, on peut considérer que deux concepts sont d'autant plus similaires qu'ils partagent plus d'information, cette information étant indiquée dans la taxonomie par le plus petit concept qui les domine tous les deux. La méthode reposant sur le décompte des arêtes mesure cela indirectement : si le chemin le plus court entre deux nœuds est tout de même long, cela signifie qu'il faut remonter haut dans la hiérarchie, jusqu'à des nœuds assez abstraits, pour trouver cet ancêtre commun. Par exemple, dans WordNet, NICKEL (*pièce de 10 cents en nickel*) et DIME (*pièce de 10 cents*) sont tous les deux dominés par COIN (*pièce*), alors que la classe la plus spécifique à laquelle appartiennent à la fois NICKEL et CREDIT-CARD (*carte de crédit*) est ASSET (*avoir*).»

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

M. Sussna choisit de combiner ces deux approches du chemin le plus court et du chemin passant par le plus petit ancêtre commun. Il mesure la distance entre deux nœuds *a* et *b* par la longueur du chemin le plus court reliant *a* et *b* au sein de la sous-hiérarchie dominée par *p*, le plus petit ancêtre commun à *a* et *b* (figure 1, *infra*). Cette approximation paraît satisfaisante même si, parfois, on ne retrouve pas le chemin le plus court : dans le cas de la figure 1, le « raccourci » antonymique qui va de *a* à *b* en passant par *c*[168] est éliminé. Ce chemin peut être composé d'arêtes de différentes natures, liens hiérarchiques d'hyponymie, relations de méronymie, d'antonymie… Reprenons l'exemple de P. Resnik déjà cité au chapitre III (3.2.1). Le chemin *a* empruntant les liens hyponymiques de COIN à ASSET et de ASSET à CREDIT-CARD est de longueur 9, tandis que le chemin *b* qui emprunte les liens hyponymiques de COIN à CURRENCY, le lien d'antonymie de CURRENCY à CREDIT et les liens hyponymiques de CREDIT à CREDIT-CARD est plus court (longueur 8). M. Sussna retient ce chemin qui est mixte mais plus court.

---

quelle que soit la source de connaissances exploitée. Plusieurs auteurs ont ainsi cherché à mesurer la parenté des sens de mots à partir de leur définition dans un dictionnaire et des mots qu'elles ont en commun. (Cowie *et al*., 1992) et (Véronis et Ide, 1990), par exemple, exploitent respectivement le ***Longman Dictionary of Contemporary English*** et le ***Collins***.

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

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

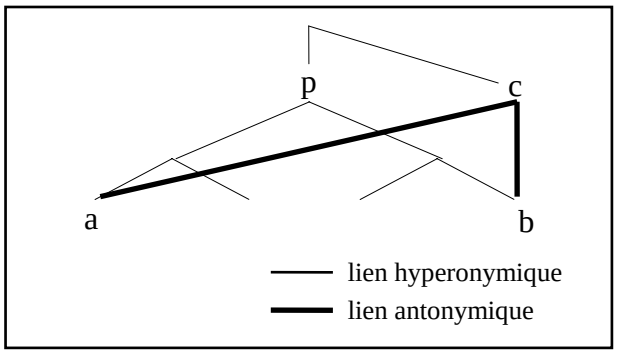


Figure 1.— Calcul du chemin le plus court au sein d'une sous-hiérarchie.

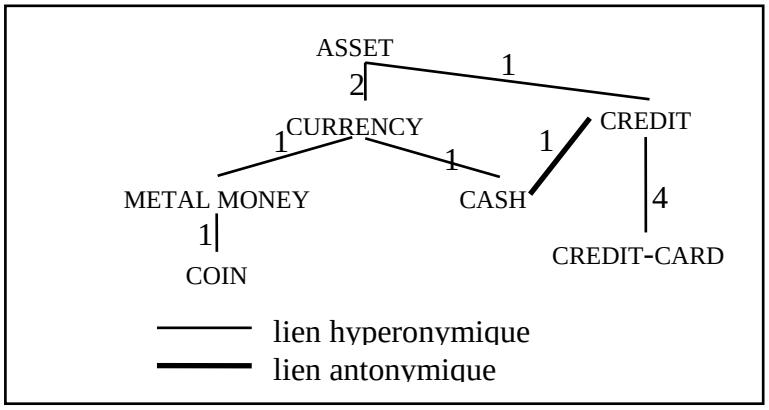


*Figure 2.— Calcul du chemin le plus court dans une sous-hiérarchie de **WordNet**. Pour aller de CREDIT-CARD à COIN, le chemin qui passe par le plus petit ancêtre commun (ASSET) est de longueur 9. Le chemin qui emprunte le lien antonymique entre CREDIT et CASH est plus court (longueur 8).*

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

- Les liens de synonymie ont un poids nul et ne comptent pas dans les mesures de distance entre nœuds : les nœuds de ***WordNet*** étant des ensembles de synonymes (*synsets*), la synonymie est une relation interne aux nœuds.
- Les liens d'antonymie ont le poids le plus fort.
- Les poids des liens hyponymiques et méronymiques varient avec la « dilution » de la relation qui est mesurée en fonction du nombre de liens de même type attachés aux nœuds concernés. Dans le cas, par exemple, de la relation A-POUR-PARTIE entre les nœuds VOITURE et PARE-BRISE, l'intuition est que cette relation reflète une parenté d'autant moins forte qu'une voiture comporte plus d'éléments (*i.e.* que plus de liens A-POUR-PARTIE partent du nœud VOITURE), mais d'autant plus forte,

à l'inverse, que les pare-brises entrent dans la composition de moins d'objets (*i.e.* que moins de liens A-POUR-PARTIE arrivent au nœud PARE-BRISE). De fait le mot *pare-brise* évoque quasi automatiquement une voiture.

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

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

C'est par l'expérimentation que M Sussna ajuste les différents paramètres de cette mesure. En ce qui concerne la diversité des liens à prendre en compte, M. Sussna montre, par exemple, en jouant sur les poids des différentes relations et en privilégiant les chemins hiérarchiques le long des liens hyponymiques, que l'on obtient de meilleurs résultats de désambiguïsation lorsqu'on exploite toute « la richesse des réseaux mixtes [comme WordNet], contenant à la fois des relations hiérarchiques et des relations non hiérarchiques » (Sussna, 1993). Les expériences menées par E. Agirre et G. Rigau (1996), qui donnent une « densité sémantique » dans ***WordNet*** comme mesure de la parenté entre les sens de mots, semblent montrer en revanche que les liens méronymiques apportent peu à la désambiguïsation[169]. Les conditions expérimentales et les mesures étant différentes, il est malheureusement difficile de comparer ces résultats.

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

24.2.3 Désambiguïser un ensemble de mots

On peut donc désambiguïser un texte en retenant pour un mot donné le sens le plus proche des sens des mots voisins. M. Sussna propose une méthode de désambiguïsation globale qui respecte la co-détermination des sens. En effet, « si on ne calcule qu'un sens à la fois, comme le font la plupart des approches numériques de la désambiguïsation de mots, la question se pose de savoir s'il faut et comment on peut tenir compte du

[169] « *A priori*, une mesure de densité calculée à partir de relations plus nombreuses devrait d'autant mieux rendre compte de la notion de parenté sémantique et on pourrait s'attendre à de meilleurs résultats [de désambiguïsation] . Les expériences […] ont montré que la différence est négligeable ; ajouter l'information méronymique n'améliore pas la précision et n'augmente la couverture[169] que de 3% environ. » (*ibid*.) Ici, la couverture correspond à la proportion de noms effectivement désambiguïsés.

fait qu'un sens a été choisi pour un mot quand on cherche à désambiguïser le mot suivant ?» (Guthrie *et al*., 1994).

M. Sussna cherche à désambiguïser non pas un mot en fonction de son contexte mais un ensemble de mots conjointement en tenant compte de leur « contrainte mutuelle » (Sussna, 1993). Cela suppose de considérer toute la combinatoire des sens possibles, de calculer une distance binaire pour chaque couple de mot et de retenir la combinaison qui minimise la distance globale (l'énergie), somme des distances binaires. Le calcul de cette contrainte devient malheureusement vite prohibitif[170]. M. Sussna propose donc de désambiguïser conjointement les premiers mots d'un texte et de poursuivre « au fil du texte » en désambiguïsant chaque mot en fonction des sens retenus pour les mots qui le précèdent. Le contexte pris en compte dans le cas général est donc le seul contexte antérieur.

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

24.3 De la désambiguïsation lexicale à la recherche documentaire

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

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

---

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

24.3.1 La granularité de la description lexicale

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

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

24.3.2 La couverture des bases lexicales

L'exploitation de bases générales pour les tâches d'indexation pose un problème de la couverture. On a vu (chapitre III.1, *supra*) que les bases lexicales générales comme ***WordNet*** ne couvrent que partiellement les corpus spécialisés. Or les systèmes de recherche documentaire doivent indexer tout type de texte, des textes spécialisés comme des articles de presse. La question de la couverture est donc cruciale[171]. R. Krovetz (1991) indique que 50 à 60 % des mots susceptibles d'être retenus comme clefs d'indexation par un système de recherche documentaire sont absents du *Longman Dictionary of Contemporary English*. E. Agirre et G. Rigau (1996), qui travaillent sur un ensemble de textes diversifiés (différents types d'articles de presse, textes scientifiques et humoristiques), signalent que 11% des noms de leur corpus ne figurent pas dans ***WordNet***. C'est donc autant de mots qui ne peuvent pas être désambiguïsés.

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

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

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

25. UN MEME PARTI PRIS D'EMPIRISME

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

25.1 Fonder une sémantique sur les corpus

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

Au cours des années 1970 et 1980, c'est surtout l'Intelligence Artificielle qui s'intéresse à l'analyse sémantique de textes[172]. L'approche retenue est celle d'une compréhension en profondeur avec l'objectif de construire une représentation logico-sémantique, de la phrase, du paragraphe ou du texte. Il s'agit de modéliser les événements et situations dont parle le texte[173]. Mais, en dépit de leur intérêt théorique, la plupart de ces travaux n'ont pas été testés en vraie grandeur sur des textes réels, de plus d'une page, portant sur des domaines variés, comportant des mots inconnus et parfois mal rédigés, etc.

De même qu'en syntaxe, les techniques d'analyse robustes ont progressivement remplacé les techniques traditionnelles dans les systèmes destinés à traiter de gros volumes de textes tout-venant, de nouvelles approches sont aujourd'hui explorées pour l'analyse sémantique. Sous l'impulsion des besoins en matière de recherche d'information ou d'aide à la lexicographie spécialisée, l'objectif s'est déplacé. On ne cherche plus à comprendre tout le texte, à le représenter dans toute sa complexité, ses implicites et ses nuances de sens. Seule une partie du texte est pertinente, la représentation cible est généralement prédéfinie et on néglige les nuances de sens, les buts du locuteur, les présupposés et implicites, etc. L'accent porte désormais sur les problèmes de structuration lexicale avec notamment la désambiguïsation sémantique des mots, le calcul de contraintes de sélection, les phénomènes de synonymie, de parenté ou de classe sémantique et plus largement le repérage des relations lexicales.

---

[172] Cf. (Herzog et Rollinger, 1991).

[173] Cela suppose tout à la fois de résoudre les anaphores, de repérer les variations de la prise en charge énonciative, de saisir la portée de telle négation ou de tel quantificateur, d'identifier les relations structurant l'ensemble du discours, etc.

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

25.2 Exploiter des résultats approximatifs

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

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

Les résultats obtenus sont parcellaires. Souvent, seuls les noms sont pris en compte. Il y a plusieurs raisons à cela. La fiabilité des analyseurs ne permet pas toujours d'exploiter les contextes verbaux. La description lexicale des noms dans un réseau comme ***WordNet*** est plus riche et plus structurée — donc plus exploitable — que pour les autres catégories. Enfin, les techniques à mettre en œuvre ou les relations à exploiter diffèrent : on ne décrit pas un adjectif ou un verbe comme on décrit un nom. Pourtant, la description lexicale des adjectifs et des verbes est importante et des verbes peuvent être de bonnes clefs d'indexation (pour les corpus spécialisés notamment). Des méthodes ont été proposées pour décrire les adjectifs ou les verbes[174], mais tout un travail d'expérimentation et de mise au point reste à faire pour construire

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

automatiquement des ébauches d'entrées de dictionnaires effectivement exploitables. Quant à la question de la désambiguïsation des verbes, R. Basili et ses collègues (1997) soulignent qu'elle est peu explorée.

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

25.3 Combiner des techniques simples

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

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

Plus généralement, il s'agit de trouver le bon dosage des méthodes linguistiques et statistiques. (Sussna, 1993) semble postuler que la description la plus riche est nécessairement la plus appropriée. Cela ne va pas de soi. Nous avons vu que des distinctions fines de sens peuvent n'être pas pertinentes pour la sélection de documents (voir *supra*, 3.3.1).

De la même manière, il n'est pas certain que la lemmatisation systématique (Church, 1995) ou la morphologie dérivationnelle, avec notamment le regroupement des mots appartenant à la même famille dérivationnelle (*stemming*) (Gaussier et coll., 1997), améliore les performances de la recherche documentaire. Par ailleurs, le travail de G. Grefenstette (1993) le montre, les traitements linguistiques sont lourds et peuvent souvent être convenablement approchés — parfois supplantés — par des techniques frustes.

25.4 Modéliser par ajustements successifs

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

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

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

Il n'est donc pas de « bon » modèle dans l'absolu. Il n'existe que des modèles opératoires qui sont utiles à l'utilisateur final dans le cadre d'une application donnée. Seul le lexicographe peut dire si les ébauches d'entrées de dictionnaires construites automatiquement lui fournissent effectivement un bon point de départ. C'est dans la mesure où la désambiguïsation lexicale telle que l'envisage M. Sussna permet d'améliorer significativement la précision de la recherche de documents qu'elle présente un intérêt, par exemple. Le verdict d'utilité est la seule

véritable évaluation possible. La maturité du domaine ne permet malheureusement pas toujours de mener cette évaluation globale à bien, mais l'exemple des entrées de dictionnaire construites par SEXTANT montre néanmoins la fécondité de cette démarche empirique.

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

25.5 Expérimenter pour mieux expliquer

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

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

Pourtant, ces expériences devraient progressivement permettre de mieux comprendre en retour les phénomènes que l'on cherche à modéliser. La diversité des conditions expérimentales fait qu'il est souvent difficile de tirer des conclusions générales sur les propriétés de telle mesure, l'importance de tel paramètre ou l'adéquation de tel modèle et nos connaissances en la matière sont encore parcellaires et fragiles. Pourtant, l'expérimentation systématique consistant à tester un à un différents paramètres comme le font M. Sussna (1993) ou G. Grefenstette (1994a), la confrontation de différentes mesures sur les mêmes données expérimentales, comme le fait (Daille, 1994) par exemple, commencent à porter leurs fruits. La convergence des résultats de différents auteurs (Sussna, 1993 ; Agirre et Rigau, 1996 ; Resnik, 1995b) montre que la parenté sémantique d'un ensemble de mots est perçue comme d'autant plus grande que leurs sens sont plus précis[175].

Le cas du score d'association est exemplaire de cette démarche empirique. K. Church et P. Hanks ont proposé (1990) de mesurer la force de cooccurrence de deux mots par une mesure fondée sur la notion d'information mutuelle et empruntée à la théorie de l'information. Ils ont

[175] Pour un sens donné, on peut mesurer ce degré de spécificité ou « *contenu informationnel* » (Resnik, 1995b) par la hauteur du nœud qui le représente dans une hiérarchie comme ***WordNet*** ou par le nombre de nœud que ce nœud domine.

montré l'intérêt et la diversité des résultats qu'elles permettait d'obtenir. À leur suite, de nombreux auteurs ont eu recours à cette mesure (Hindle, 1990 ; Resnik, 1995b). Pourtant le choix de cette mesure n'est jamais réellement justifié : on en explicite les propriétés formelles, mais sans expliquer pourquoi cette mesure est pertinente pour mesurer des contraintes de sélection. La convergence de différentes expériences montre cependant qu'en donnant un poids important aux événements rares et en soulignant les emplois « spécialisés »[176], le score de cooccurrence fait ressortir les expressions figées, ce qui est précieux dans une perspective lexicographique : l'association de *œil* et de *boeuf*, dans *oeil de bœuf*, est intéressante pour la description du mot *boeuf*. Mais ceci explique à l'inverse que cette mesure soit mal adaptée à la modélisation conceptuelle d'un domaine, ce que (Habert *et al.*, 1996) met en évidence. Pour décrire le concept auquel renvoie un mot, ses propriétés et les relations dans lequel il entre, il faut au contraire éliminer les attirances proprement lexicales et s'appuyer davantage sur les associations banales comme *manger/élever du bœuf*, *viande de bœuf*, *bœuf cuit*, etc. L'information mutuelle est donc un bon indice lexicographique mais un mauvais outil de modélisation conceptuelle. Par ailleurs, cette mesure qui « met l'accent sur les phénomènes rares » (Basili *et al*., 1993b, p. 179) est peu adaptée aux contextes syntaxiques : elle serait utile « si on pouvait se fier entièrement aux analyses » (*ibid*.), mais elle donne en fait trop d'importance à des relations « dues à des ambiguïtés syntaxiques ou des erreurs d'analyse » (*ibid*.). C'est la multiplication et la confrontation des expériences utilisant la mesure de l'information mutuelle et la comparaison avec des mesures différentes qui permet de tirer des conclusions de portée un peu générale, de progressivement mieux comprendre ses propriétés comme mesure de distance entre les mots et de cerner les conditions de son utilisation.

[176] Le fait pour un mot de figurer toujours ou très souvent dans le(s) même(s) contexte(s).

CHAPITRE V

LE LANGAGE AU FIL DU TEMPS : CORPUS ET DIACHRONIE

26. DEFINITIONS ET ENJEUX

L'écoulement du temps structure de nombreux corpus, sans qu'ils permettent pour autant la saisie de l'évolution du langage. La volonté de créer des dictionnaires reposant sur l'usage effectif et son changement a par exemple contribué à la création de corpus électroniques intégrant des données de différentes périodes. C'est le cas du *Trésor de la Langue Française* (INaLF, CNRS) qui s'appuie sur une base de textes de plus de 160 millions de mots, s'étalant du XVI^{e} au XX^{e} siècle. Toutefois, de tels corpus ne constituent pas forcément des corpus adaptés aux études diachroniques. Le registre littéraire y domine, au détriment d'autres registres. La dimension temporelle structure également d'autres corpus, encore plus spécialisés. Corpus mono-émetteur : c'est le cas de ***Mitterrand1***, dont les textes s'égrènent sur le premier septennat. Corpus pluri-locuteurs : c'est le cas des résolutions générales des quatre grandes confédérations syndicales ouvrières françaises étudiées entre 1971 et 1976 (Bergounioux *et al.*, 1982)[177]. Ces corpus sont de la même manière restreints à un registre (ou à des variations sur un même registre) : entretien, interview et discours de circonstance pour ***Mitterrand1***, résolutions de congrès pour (Bergounioux *et al.*, 1982). Le temps intervient, mais on ne peut saisir son rôle que sous un angle limité : une thématique, un domaine, ou un genre bien défini.

A côté de ces corpus de fait spécialisés, se constituent des corpus « historiques ». Ils sont destinés explicitement à l'étude de l'évolution de la

[177] Le chapitre IX aborde la mesure de l'évolution lexicale de tels corpus.

langue. Nous présentons en détail un corpus de ce type : ***Archer***, en section 2, ainsi que les problèmes de représentativité et de constitution de tels corpus. L'évolution de la langue peut être examinée sur la courte durée, sur le moyen terme, ou sur le long terme. Nous rendons compte d'études relevant de ces différentes temporalités en section 3. Nous abordons enfin en section 4 les problèmes méthodologiques propres aux corpus historiques.

27. UN CORPUS POUR L'ETUDE DE LA DIACHRONIE : *ARCHER*

Les analyses diachroniques de l'anglais disposent du corpus d'***Helsinki*** d'1,5 millions de mots (Kytö, 1993b). La période couverte va de 750 à 1700[178]. Le corpus ***Archer***[179] (Biber *et al.*, 1994) complète la tranche chronologique couverte[180]. D. Biber, E. Finegan et D. Atkinson (1994, p. 7-13) montrent les usages possibles d'un tel corpus historique. Ils utilisent par exemple la distinction établie par Biber (cf. chapitre I) entre production informationnelle (qui favorise noms, prépositions, adjectifs attributs etc.) et production « impliquée » (qui privilégie le présent, l'omission de *that*, les contractions, les démonstratifs, la première personne, le pronom *it*, *BE* comme verbe principal, les pronoms indéfinis, etc.). Si l'on compare les registres, théâtre, lettres et journaux intimes se font plus « impliqués » depuis le XVII^{e} siècle, tandis que la médecine devient plus « informationnelle ». La comparaison entre anglais et américain sur la même durée montre que les registres américains sont généralement plus informationnels que leurs équivalents anglais.

27.1 L'anglais et l'américain de 1650 à aujourd'hui

Archer[181] a été constitué pour permettre l'étude diachronique de l'anglais et de l'américain entre 1650 et aujourd'hui par le biais de dix « registres », qui mêlent thématiques et genres[182]. Les registres sont les suivants pour l'écrit : journaux intimes, lettres, fiction, écrits journalistiques, médecine (anglais seulement), science (anglais seulement), décisions de justice (américain seulement), et pour l'écrit lié à l'oral (c'est-à-dire imitant l'oral ou servant de base à une production orale) : les conversations fictives, le théâtre, les sermons et homélies.

---

[178] Des documents écossais (1450-1700) et américains (1600-1700) constituent deux corpus complémentaires (Kytö, 1993).
[179] A Representative Corpus of Historical English Registers.
[180] Il y a donc recouvrement pour la période 1650-1700, ce qui autorise des comparaisons fructueuses sur les choix faits pour représenter ce laps de temps (cf. *infra*).
[181] Le corpus rassemblé à Cambridge (English Faculty) pour la période 1600-1800 s'inscrit dans la même perspective (Wright, 1993).
[182] Dans la même acception qu'au chapitre I.

Archer est organisé par périodes de cinquante ans pour que l'on puisse examiner l'évolution, les flux et les stabilités sur des périodes relativement courtes. L'américain n'est dans l'immédiat représenté que par trois périodes : deuxième moitié des XVIII^{e}, XIX^{e} et XX^{e} siècles. L'anglais l'est pour les neuf périodes. Pour chaque période de cinquante ans et chaque registre, un échantillon de 20 000 mots[183] est constitué. ***Archer*** totalise 1,7 million de mots.

27.2 Echantillonnage des registres

Le choix de textes relevant des registres visés se heurte à plusieurs obstacles. En premier lieu, les ressources bibliographiques sont organisées thématiquement et non par registres. Ainsi, une des sources bibliographiques consultées, à l'entrée *lettres*, renvoie en fait aux manuels d'écriture de lettres, ce qui ne correspond pas à l'objectif visé : la correspondance privée authentique. En second lieu, « les distinctions de registre d'une période peuvent ne pas correspondre exactement avec celles d'une autre période[184]. Les registres ne restent pas nécessairement distincts l'un de l'autre au fil du temps. Bien sûr, les registres émergent à un moment donné de l'histoire, pas nécessairement tous au début d'une période d'investigation ni au début d'une période de cinquante ans retenue » (Biber *et al.*, 1994, p. 5).

M. Kytö (1993) témoigne de la complexité des paramètres à prendre en compte pour rassembler des données représentatives de l'américain entre 1600 et 1700 , dans le cadre d'un autre corpus historique. Seuls sont retenus les documents écrits (et éventuellement imprimés) aux Etats-Unis, et pour la période commençant en 1670, date qui sépare la première génération d'immigrants de ses descendants, provenant d'auteurs nés dans ces colonies (ou établis depuis suffisamment longtemps). Les dates d'installation différentes des colonies du Sud (Virginie, premières arrivées en 1607) et du Nord (Plymouth, 1620, baie de Massachusetts 1630, etc.) amènent à constituer des échantillons distincts pour rendre compte de leurs histoires langagières propres. Certains registres caractéristiques des colonies ont été intégrés : récits de captivité, témoignages, etc. L'appréhension de l'oral ne peut s'effectuer que par des biais : « Le langage de tous les jours trouvé dans la correspondance privée, certains journaux intimes ou des textes faits pour être dits fournissent un moyen d'approcher le langage parlé du passé, le vrai cœur du changement linguistique. De la même manière, les écrits des immigrants les moins éduqués, qui n'auraient peut-être pas pris la plume dans leur pays d'origine mais qui étaient forcés de le faire dans les colonies, peuvent aussi nous donner des aperçus [*glimpses*] de la langue

[183] 10 fragments de 2 000 mots, pour diminuer le poids des idiolectes.
[184] Par exemple la correspondance peut relever de la littérature, voire de la philosophie, comme de l'échange purement privé aux XVI^{e} et XVIII^{e} siècles (Wright, 1993, p. 26). Finegan et Biber (1995, p. 249) expliquent l'incohérence relative de leurs résultats concernant les lettres par l'hétérogénéité de ce registre.

parlée » (*ibid.* p. 5)[185].

Pour ***Archer***, au sein d'un registre, le choix des ouvrages repose sur une procédure aléatoire[186] (au sens probabiliste)[187]. Un protocole bien défini permet également, pour chaque registre, d'extraire des fragments (pas forcément continus) de 2 000 mots[188].

27.3 Structuration temporelle

L'échelonnement des documents retenus peut avoir comme logique une périodisation. C'est le choix d'***Archer***, qui distingue donc des périodes de cinquante ans : ce sont les blocs qui sont soumis ensuite à l'analyse linguistique et statistique.

Le parti pris du corpus couvrant l'anglais de 1600 à 1800, à Cambridge (Wright, 1993), est tout autre : un étalement continu des documents, avec une ossature formée de textes sélectionnés à dix ans d'intervalle. L'objectif est ici de permettre au chercheur de choisir les intervalles qui lui paraissent pertinents et de ne pas l'enfermer dans une périodisation qui peut s'avérer non valide pour sa recherche.

27.4 Représenter les états de langue ou des idiolectes ?

De quels usages les corpus historiques constitués sont-ils représentatifs ? Une des réponses possibles est celle qui sous-tend la création d'***Archer*** : les variations observées relèvent des genres ou des types textuels sous-jacents. Si l'on veut étudier l'évolution d'une langue, il faut articuler l'échelonnement des textes dans le temps avec leur stratification en genres qui ont une cohérence et un mouvement propres. D'où une démarche d'échantillonnage aléatoire, utilisant des extraits courts, mais nombreux. Finegan et Biber (1995, p. 252) soulignent ainsi que la représentation du genre *sermons* est probablement plus satisfaisante dans ***Archer*** que dans ***Helsinki***, même si ce dernier corpus comprend des textes entiers qui totalisent un nombre de mots plus important. ***Helsinki*** en effet utilise les sermons de deux prêcheurs seulement, tandis

[185] Cet article fournit des extraits significatifs de tels documents (*ibid*., p. 5-8).

[186] Ainsi, pour la fiction anglaise, le répertoire *Oxford Companion to English Literature* (*OCEL*) a été utilisé. Les 1 099 pages de l'*OCEL* ont été divisées par le produit du nombre de périodes et de textes requis pour chaque période, ce qui a fourni un intervalle de 13 pages. Le numéro de la première page considérée a été tiré au hasard, puis on a examiné la page suivante à 13 pages d'intervalles et ainsi de suite. Pour les textes de fiction par exemple, sur chaque page examinée, on a pris le premier auteur anglais ayant écrit un roman dans une des périodes retenues et on a choisi son 3e roman s'il y en avait 3 ou plus (ou son 2e ou son unique roman). On a continué jusqu'obtenir le nombre de textes nécessaires pour toutes les périodes (ce qui a nécessité plusieurs passes sur l'*OCEL*, en tirant à chaque fois un nouveau numéro au hasard pour la première page considérée).

[187] On reviendra au chapitre IX sur les raisons de ce choix.

[188] Par exemple, pour les textes journalistiques ou scientifiques anciens, les documents ont souvent une taille inférieure à 2 000 mots. Il faut alors regrouper. Inversement, dans les périodes récentes, la longueur des textes oblige à prélever les 500 premiers et derniers mots, ainsi qu'un empan de 1 000 mots au milieu, pour ne pas sur-représenter certains « sites » particuliers des textes (introduction, conclusion, etc.).

qu'un échantillon plus élevé de prêcheurs figure dans ***Archer***.

D'autres travaux (Wright, 1993, p. 27-29) insistent au contraire sur la dimension idiolectale des observations. S. Wright (*ibid.* p. 28) cite par exemple les recherches sur l'emploi de certains marqueurs relatifs : « [...] au début du XVII^{e} siècle, le système des relatives différait du système actuel en ce que le pronom *which* pouvait optionnellement servir à renvoyer à un antécédent humain aussi bien qu'à un antécédent non humain. Cependant, progressivement, c'est le pronom *who* (à la place de *which*) qui a été choisi pour renvoyer à des antécédents humains. Hope (1990) a montré que le choix des marqueurs relatifs dans les œuvres de Shakespeare et Fletcher était basés sur deux systèmes en compétition. Alors que celles de Fletcher sont typiques de l'association moderne entre le relatif *who* et des antécédents humains, l'usage suivi par Shakespeare suggère que ce trait n'est pas un facteur aussi significatif pour son choix. Pour ces deux écrivains donc, la sémantique du système de marqueurs relatifs a des valeurs différentes. » Le rassemblement de données textuelles plus importantes pour un groupe d'auteurs contemporains a pour objectif alors de caractériser l'usage commun de ce groupe par rapport aux idiolectes de chacun des auteurs[189]. Se pose aussi la question de la part de la manipulation stylistique de la langue, de l'idiolecte et de l'usage du moment.

28. ÉTUDES DE LA DIACHRONIE

Les corpus électroniques permettent d'examiner l'évolution de certains phénomènes langagiers sur de très courtes durées (d'une année sur l'autre, par exemple), sur le moyen terme (quelques décennies) et sur le long terme : on peut alors comparer des états de langue reconnus comme distincts dans la tradition linguistique (ancien français / moyen français / français classique / français moderne) ou examiner les changements au fil des siècles.

28.1 La courte durée

J. Sinclair a forgé le terme de « corpus de suivi » (*monitor corpus*) pour désigner des flux continus de textes permettant l'analyse chronologique, année par année par exemple, de données langagières. Cette notion était au départ une vue de l'esprit. De plus en plus de textes sont désormais directement sous forme électronique. C'est le cas de quotidiens employant une langue « tenue » comme *Le Monde*, *The Guardian*, édités sous forme de CD-ROM. C'est le cas aussi des bandes de

[189] Voir (Wright, 1993, p. 30-34) pour une discussion du statut à donner aux emplois par Joseph Addison des différentes formes de relatives. S. Wright prend nettement le contrepied de l'interprétation que fournissent Biber et Finegan des mêmes faits.

photocomposition de journaux mises à disposition des créateurs de corpus. On peut donc comparer les ensembles constitués pour chaque année, ou examiner les apports d'une année donnée[190].

A. Renouf (1993) détaille l'utilisation en ce sens du *Times*, de novembre 1990 à septembre 1991. Un premier filtrage isole les mots nouveaux, en les répartissant en noms propres, acronymes et mots « ordinaires ». Le classement de ces derniers renseigne sur les mécanismes à l'œuvre et leur productivité relative : formations à base d'onomatopées, jeux de mots, mots-valises, « composés », doublons dérivatifs (*indifferentness*), suffixations (*eco-terrorism*, *executivedom*), préfixations (*euroconvertible*) et conversions, etc. Par exemple, *gate*, par analogie avec *Watergate*, n'est guère productif en mars 1991 (seul ce mot est utilisé) mais donne naissance en fin 1992 à *iraq(-)gate*, *dianagate*, *camillagate*, *threshergate*. A. Renouf (*ibid.*, p. 286-287) donne aussi les 50 préfixes (*non-*, *re-*, *over-*, etc.) et suffixes (*-like*, *-based*, *-style*, etc.) les plus fréquents dans les composés de mars 1991.

28.2 Le moyen terme

La constitution des premiers corpus de référence pour l'anglais remonte aux années soixante, avec ***Brown*** et ***LOB***. Ces deux corpus fournissent un échantillon voulu représentatif de l'usage, américain d'un côté, anglais de l'autre, en 1961 précisément, au sein d'un certain nombre de registres. Plus de trente ans nous séparent de ces « instantanés » du début des années soixante. Aussi peut-on s'en servir pour examiner les écarts avec l'usage actuel.

C'est l'objectif de C. Mair (1995). Il compare l'emploi de *help* dans ***Brown*** et ***LOB*** avec l'usage en 1991. C. Mair a constitué pour ce faire un corpus selon les mêmes critères que ***LOB***, à ceci près que les textes retenus sont de 1991. Il appuie également son analyse sur le CD-ROM du journal *The Guardian* pour la même année. Il examine l'évolution des constructions suivantes de *help* :

> + to infinitif (*Maybe he will help to turn our fair city into a 'ghost town'*)
> + infinitif seul, éventuellement précédé d'un SN sujet « logique » de cet infinitif (*I helped him mend his bicycle*)

La deuxième construction est généralement présentée comme un américanisme dans les grammaires anglaises. Une étude détaillée indique que la première est effectivement la variante dominante en anglais dans les années soixante. Le corpus de 1991 montre (*ibid.*, p. 264) d'une part que la fréquence de *help* avec un complément infinitif s'accroît sensiblement par rapport à 1961 et d'autre part que la construction avec

[190] Pour les corpus de suivi, le problème n'est pas de réaliser une édition électronique « propre », exempte de coquilles, faisant autorité, mais de pouvoir utiliser au plus vite des données vastes qui vont se trouver rapidement remplacées par d'autres (Blackwell, 1993, p. 101). Le nettoyage ne vise pas la perfection. Il doit simplement permettre le fonctionnement des outils logiciels d'exploration des données. Vu la taille des données traitées, il doit être entièrement automatique ou limiter au maximum l'intervention humaine.

infinitif seul domine désormais (en particulier sans SN sujet « logique » de l'infinitif). La construction avec infinitif seul domine également dans le CD-ROM de 1991 du journal *The Guardian*. Comme il s'agit d'un journal dont la langue est « tenue », cette prédominance montre que la construction en cause a perdu la connotation de « relâchement » qui était la sienne trente ans auparavant. C. Mair voit dans cette évolution l'indice d'une « grammaticalisation », définie comme la transformation au fil du temps de certaines formes lexicales en simples marques grammaticales. *Help* se viderait progressivement de son sens et deviendrait un simple « étai » pour l'infinitif associé[191]. Pour C. Mair (*ibid.*, p. 267), en outre, l'opposition faite par les grammaires entre les deux constructions n'est pas tout à fait exacte. L'anglais et l'américain suivraient un mouvement parallèle, quoique décalé, dans l'évolution de l'utilisation de *help*.

28.3 La longue durée

28.3.1 La position des adjectifs en moyen anglais tardif

H. Raumolin-Brunberg (1994) étudie la position des adjectifs en moyen anglais tardif (1350-1500). Elle s'appuie sur les données d'***Helsinki***. Elle examine particulièrement l'hypothèse avancée par plusieurs chercheurs selon laquelle la position de base serait post-nominale : on trouverait globalement plus d'adjectifs après qu'avant le nom ; pour les adjectifs pouvant se présenter dans les deux positions, la post-position serait plus fréquente ; enfin, la position après le nom serait non marquée. H. Raumolin-Brunberg limite son étude à la prose pour que n'interviennent pas les contraintes sur l'ordre des mots propres à la poésie. Le sous-corpus examiné comprend 200 000 mots.

Les constats effectués dans ***Helsinki*** contrecarrent nettement l'hypothèse formulée ci-dessus. La comparaison de deux sous-périodes (1350-1420 et 1420-1500) ne montre pas d'évolution sur la position de l'adjectif, là encore contrairement à certaines propositions. En outre, l'écart entre les proportions pour les occurrences et les lemmes indique que beaucoup des adjectifs précédant habituellement le nom sont très fréquents (*great*, *good*, *holy*, etc.). Les post-posés sont au contraire peu fréquents[192], comme le montre le tableau suivant :

place de l'adjectif	occurrences	%	formes	%
pré-nominal	5 197	92,3	531	73,1
post-nominal	432	7,7	195	26,9
total	5 629	100	726	100

[191] Un peu comme dans les constructions à verbe support du type *prendre peur* où le nom véhicule l'essentiel du sémantisme, le verbe apportant des indications temporelles et aspectuelles.
[192] Les adjectifs qui apparaissent uniquement post-posés sont à 90 % d'origine latine ou française.

Les résultats obtenus sont également très proches d'études faites pour l'anglais contemporain. Enfin, l'examen des divers registres représentés dans le sous-corpus ne manifeste pas d'écarts significatifs dans le placement des adjectifs par rapport aux constats globaux qui viennent d'être donnés.

Au regard de ces résultats, H. Raumolin-Brunberg conclut à la primauté de la position antéposée de l'adjectif en anglais, tout au long de son histoire.

28.3.2 L'alternance *that* / zéro

En anglais, après certains verbes comme *hear*, *hope*, *know*, *think*, *say* et *tell*, certaines propositions objet peuvent être introduites par *that* (*I hope that becoming a catholic will give you peace of mind*) ou rester non marquées *(I told him I had a letter from you*). Cette alternance et ses conditions ont largement été étudiées. Les données d'***Helsinki*** ont permis de montrer une tendance générale à la progression de la construction zéro entre 1350 et 1710.

Finegan et Biber (1995) reprennent l'étude de cette alternance en utilisant ***Archer***, sur la période allant de 1650 à 1990. Mais ils se restreignent à trois genres : les lettres, les sermons et les articles médicaux. Toutes périodes confondues, la répartition par construction et par registre est la suivante :

	that	zéro
sermons	89 %	11 %
médecine	83 %	17 %
lettres	53 %	47 %

Paradoxalement, les résultats pour les articles médicaux et les sermons vont à contrecourant de la tendance mise en évidence pour ***Helsinki***[193]. Au contraire, ces deux registres favorisent continûment et de plus en plus nettement la construction avec *that* par rapport à la construction zéro. Finegan et Biber interprètent ce décalage par une progression plus générale de ces registres vers une forme plus cultivée (*literate*) et moins orale. Les lettres témoignent d'une évolution comparable, mais plus atténuée (avec un étonnant renversement de tendance pour la période 1900-1949, où la construction zéro domine).

Ces évolutions décalées poussent à multiplier les points de vue dans l'analyse globale de changements linguistiques. Finegan et Biber examinent d'ailleurs les attirances de certains des verbes majeurs pour chacune de ces deux constructions, toutes périodes confondues : « [...]

[193] Finegan et Biber (*ibid.*, p. 251-253) montrent dans le détail les difficultés d'une comparaison des résultats sur ***Helsinki*** et sur ***Archer*** pour la période approximativement partagée par ces deux corpus (1640-1710 et 1650-1699 respectivement). Les principes d'échantillonnage diffèrent, on l'a vu. La taille réduite des parties correspondant à cette période pour les deux corpus fait aussi obstacle.

les verbes *say*, *tell* et *know* montrent une forte préférence pour *that* dans les trois registres, tandis que *think* montre une préférence nette pour la construction zéro, du moins en médecine et dans les lettres » (*ibid.*, p. 250).

28.3.3 L'évolution des démonstratifs en français

En français, les démonstratifs ont connu un changement morphologique radical. Aux XI^e^ et XII^e^ siècles, s'opposent sémantiquement deux paradigmes de démonstratifs. Le premier (désormais CIST) est issu du latin vulgaire *ecce iste*, le second (désormais CIL), d'*ecce ille*. Le premier exprime la proximité, le second l'éloignement, temporel ou spatial, soit par rapport à l'auteur, soit par rapport à l'un des personnages[194]. Chacune des formes peut être aussi bien pronom (*Cil vient*) que déterminant (*Cil chevaliers vient*) et il existe en outre des formes longues préfixées par *i-* : *icelui,* etc. Rappelons que l'ancien français possède une déclinaison opposant deux cas : le cas-sujet (issu du nominatif latin) et le cas-régime (issu de l'accusatif latin). S'ajoute parfois, c'est le cas pour les démonstratifs, un second cas-régime singulier (issu du datif latin). Au total, 14 formes différentes (28 si l'on inclut celles préfixées en *i-*). À partir du XVII^e^ siècle, le paradigme des pronoms (*Celui-ci vient*) est totalement séparé de celui des déterminants (*Cet homme vient*). Une étape marque le passage d'un système à l'autre. Au XII^e^ siècle, apparaît au nord de la France une nouvelle forme de cas-régime masculin pluriel : *ces*, toujours déterminant, va ensuite être employé également au féminin pluriel. Fin XII^e^-début XIII^e^ siècle, apparaît *ce*, déterminant masculin singulier au cas-régime, employé uniquement devant un mot commençant par une consonne (*ce chevalier*). C'est en fait un nouveau paradigme qui émerge, le troisième : *ce* / *ces*, uniquement déterminant et toujours atone, sans opposition de genre au pluriel, et sémantiquement indifférencié (pas d'opposition proximité / éloignement).

Ce changement profond n'a pas d'équivalent dans la plupart des autres langues romanes, où les formes de démonstratifs continuent à être employées à la fois comme déterminants et comme pronoms. Il reste énigmatique : les changements phonétiques ne suffisent à expliquer ni la spécialisation globale des paradigmes ni la sélection des formes survivantes au sein de chaque paradigme.

L'objectif de C. Marchello-Nizia (1995, p. 115-181) est d'expliquer dans le détail la répartition et l'évolution des différentes formes. Les hypothèses qu'elle propose s'appuient sur des constats que seul permet le traitement de très gros corpus[195]. Elle souligne en effet (*ibid.*, p. 138-139) : « Par généralisation ou simplification abusive, on gomme le fait que ce n'est pas

[194] L'opposition sémantique entre les deux séries, indéniable, est plus complexe. Elle a suscité de nombreuses analyses (Marchello-Nizia, p. 129-130). L'hypothèse actuellement la plus satisfaisante, selon C. Marchello-Nizia, est celle de G. Kleiber (*ibid*., p. 129-137). Pour ce dernier, les formes en CIST indiquent au destinataire qu'il faut opérer l'appariement référentiel à partir du contexte d'énonciation immédiat de l'occurrence (contexte spatio-temporel représenté ou contexte énonciatif ou discursif), ce qui n'est pas le cas pour les formes en CIL.

[195] Cf. section 4.1 sur la taille des corpus historiques.

tout le paradigme de CIL qui est devenu pur pronom, mais seulement quatre formes sur sept : *celui*, *celle*, *ceux*, *celles* ; *cil*, *cel* et *celi* ont disparu. Pour *cil*, on peut dire qu'il s'agissait d'une forme de cas-sujet (singulier ou pluriel), et dès lors que la déclinaison disparaissait, les formes qui instanciaient les différents cas devaient disparaître. Mais pourquoi est-ce *celui* qui s'est conservé et non *cel*, et pourquoi à l'inverse pour le féminin est-ce *celle* et non pas *celi* qui s'est conservé ? De même, ce n'est pas tout le paradigme de CIST qui s'est conservé en devenant pur déterminant. Sur six, seules deux formes, la forme du féminin singulier *cette*, et celle du masculin singulier devant voyelle *cet*, viennent directement du paradigme CIST. *Ce* n'en provient pas, non plus que proprement le pluriel épicène[196] *ces*[197]. Les autres formes, au nombre de quatre (*cist*, *cestui*, *cez*, *cestes*), ont disparu. *»*

C. Marchello-Nizia s'appuie sur un important corpus d'ancien et de moyen français. Pour l'ancien français, ont été utilisés seize textes en vers ou en prose (*ibid.* p. 147-148), soit près de 685 000 mots, s'échelonnant de 1100 environ à 1300 environ. Ces textes se situent dans le domaine littéraire, central dans les recherches des médiévistes, et une concordance est disponible pour chacun d'eux. Ils comprennent 8 237 démonstratifs. Pour le moyen français (XIV^{e} et XV^{e} siècles), le corpus utilisé pour la constitution du *Dictionnaire du Moyen Français* (INaLF, Nancy), qui compte environ 4 millions de mots et qui est d'origine plus variée[198], a fourni près de 36 000 occurrences de démonstratifs.

L'examen détaillé des concordances des formes longues (préfixées en *i-*, suffixées en *-ui* / *-i*, ou portant les deux affixes) dans le corpus d'ancien français[199] permet de mieux cerner les notions de *«* soulignement *»*, d'*«* expressivité *»*, de *«* renforcement *»*, utilisées jusqu'alors. Ces formes sont en effet employées en début de phrase ou de vers. Elles sont pronoms dans 3 cas sur 4 pour les formes suffixées en *-ui* / *i* et déterminants dans deux tiers des cas pour les formes préfixées en *-i* . Elles déterminent alors le plus souvent un substantif complément d'objet placé en tête de phrase. Elles mettent en évidence cette construction, marquée à cette époque.

A partir de ces observations, C. Marchello-Nizia (*ibid.*, p. 144) formule l'hypothèse d'une répartition des démonstratifs en trois groupes : les formes toujours atones (*ces* et *ce*), les formes toujours toniques (les formes longues) et les formes pouvant être atones ou toniques (*cil*, *cel*, *cele*, *ceus* et *cist*, *cest*, *ceste*). C'est dépasser l'opposition déterminant / pronom et prendre en compte la dimension accentuelle.

Les cas-sujets masculins singuliers *cil* et *cist* suivent bizarrement une évolution décalée : *cist* s'efface à partir de 1250, en lien avec la chute de la déclinaison, tandis que *cil* reste employé jusqu'à la moitié du XV^{e} siècle, où il connaît une disparition brutale. C'est un parallélisme avec le pronom personnel *il* qui expliquerait cette évolution de *cil* : on constate en effet

[196] Utilisable au féminin et au masculin.
[197] Cette forme provient à la fois de *cez* (de la série CIST), par évolution phonétique de l'occluso-constrictive finale [ts] en [s] et de *cels* (de la série CIL), employé de façon inaccentuée et proclitique comme déterminant. *Ce* est fait par analogie sur *ces*.
[198] 182 œuvres différentes, de longueur inégale et de divers genres (chroniques, romans, chansons de geste, poésie lyrique ou didactique, chartes, traités philosophiques, etc.).
[199] 1 027 occurrences sur 8 237 démonstratifs.

une évolution parallèle de *il* et de *cil* (*ibid.* p. 164). En outre, les comptages opérés montrent qu'en moyen français, les deux paradigmes CIST et CIL ne sont pas encore spécialisés, l'un pour les déterminants, l'autre pour les pronoms. Les emplois pronominaux sont occupés essentiellement par trois formes : *celui*, *celle*, et *cestui*. Ce serait là encore l'influence du système pronominal qui aurait joué. Ont en effet été conservées comme pronoms démonstratifs les formes (*celui*, *ceux*, *celle*, *celles*) ressemblant aux pronoms personnels employés de manière autonome (*lui*, *eux*, *elle*, *elles*), celles sans correspondant pronominal disparaissant (comme *celi*, *cesti*, *cestui*). Par ailleurs, les formes longues se spécialisent en moyen français dans la fonction de pronom, alors que dans la période précédente, la détermination focalisante les caractérisait. Ce serait aussi le contrecoup du remplacement progressif de l'accent tonique de mot à valeur distinctive, encore présent en ancien voire en moyen français, par l'accent en fin de groupe syntaxique, la détermination marquée trouvant dans *-ci* et *-là* post-fixés le moyen de souligner cet accent de groupe. Cette évolution est une deuxième étape dans le mouvement de distinction entre la catégorie du pronom et celle du déterminant, mouvement amorcé avec l'apparition du déterminant *ce* / *ces*, et achevé à la fin du moyen-âge par l'institution de formes purement pronoms.

29. PROBLEMES METHODOLOGIQUES

La constitution et l'annotation de corpus diachronique rencontrent des obstacles spécifiques. Les ressources résultantes permettent néanmoins de vérifier, de préciser les évolutions et de renouveler les explications qui en sont fournies.

29.1 Des corpus « petits » et peu annotés

La constitution même des corpus pose des problèmes spécifiques pour les états anciens d'une langue où les sources sont des manuscrits (l'ancien français par exemple). Les variantes graphiques d'une même forme peuvent être nombreuses[200]. Mais il est désormais possible de mémoriser et de relier différents types de documents. C'est le cas du projet *Charrette* dirigé par K. Uitti (Université de Princeton) : les transcriptions diplomatiques des huit manuscrits du XIIIe siècle du *Chevalier de la Charrette* de Chrétien de Troyes, soit près de 36 000 lignes pour un poème d'environ 7 100 lignes, sont reliées à une version électronique de l'édition Foulet-Uitti et aux images de ces manuscrits. La philologie voit ainsi s'ouvrir de nouvelles perspectives.

[200] Les 28 formes de démonstratifs repertoriées par C. Marchello-Nizia (1995) se réalisent en plus de 80 graphies.

Nous l'avons vu, le développement des corpus électroniques a très largement bénéficié cette dernière décennie des apports, techniques et financiers, de la communauté du TALN qui voit là une étape indispensable pour la mise au point de systèmes de traitement du langage robustes. L'accent est bien sûr mis sur la langue contemporaine. Autrement dit, il n'y a pas vraiment de raisons que beaucoup de temps et d'énergie soit consacré à la recherche sur les états de langue anciens. On peut donc escompter un retard sensible dans les techniques et les moyens mis en œuvre pour l'annotation des corpus historiques. Les corpus historiques actuels sont d'ailleurs très sensiblement plus petits que les corpus synchroniques (Finegan et Biber, 1995). Que l'on compare le million et demi de mots d'***Helsinki*** ou d'***Archer*** avec les 100 millions de mots (étiquetés, au surplus) de ***BNC***.

En dehors de ces projets de corpus conçus pour étudier la diachronie, parce qu'il est coûteux de constituer des corpus bien répartis sur les genres et les périodes, les constats sont souvent établis sur les ensembles de textes qui sont effectivement disponibles sous forme électronique mais qui ne forment pas vraiment un corpus historique au sens d'***Archer*** par exemple. Cette situation biaise évidemment les observations et leur interprétation, sans que les chercheurs qui ont recours à ces rassemblements de circonstance en soient toujours conscients.

L'annotation de ces corpus se heurte en outre à des obstacles spécifiques. Une langue à cas comme l'ancien français connait une variation importante dans l'ordre des mots, alors que les étiqueteurs et parseurs disponibles ont été conçus pour des langues où l'ordre des mots est notablement plus contraint. La connaissance du lexique et de la syntaxe de ces états de langue n'offre pas non plus le même appui à une automatisation. À l'inverse, ces corpus historiques étant destinés, pour leur très grande majorité, à rester « nus », ils ne permettent pas facilement de valider ou d'invalider des hypothèses linguistiques. Ils supposent une analyse très souvent manuelle des données[201] pour trier les faits et proposer des hypothèses, mais aussi pour comparer la représentation formelle postulée avec le corpus. Ainsi, T. Nevalainen (1994), pour étudier l'évolution de l'opposition en anglais entre les formes des adverbes en *-ly* et sans suffixe (*slowly* / *slow*) en contrastant la période 1350-1420 avec la période 1640-1710, commence par extraire d'***Helsinki*** les formes se terminant en *-ly* (elle répertorie 14 variantes graphiques du suffixe), élimine celles qui ne sont pas des adverbes ainsi que les adverbes faits sur une base nominale (*namely*), et cherche les adjectifs ayant servi de base aux adverbes ainsi isolés. Ce sont encore de simples concordances qui sont employées par Finegan et Biber (1995, p. 245) dans leur étude de l'alternance *that* / zéro après certains verbes.

[201] Même si des environnements informatiques adéquats allègent parfois la charge.

29.2 Vérifier et préciser les évolutions

C. Mair (1995, p. 260) résume assez bien ce que la linguistique diachronique va gagner dans ces nouvelles études : *«* L'approche du changement linguistique basée sur les corpus corrigera des distorsions évidentes dans la littérature actuelle sur le sujet. Il sera possible de séparer l'usuel et le normal de l'exceptionnel. À la différence de l'observateur qui enregistre l'exemple unique d'une nouvelle construction tout en omettant de noter les preuves massives de la persistance de l'ancienne construction, l'analyste de corpus sera en position de décrire les tendances statistiques avec précision. *»*

Ce constat se vérifie déjà pour l'exemple des démonstratifs en français. Les textes de la période effectivement disponibles sous forme électronique ne couvrent pas, loin s'en faut, tout ce qui est répertorié. Les conclusions et décomptes actuels seront donc sans doute infléchis[202]. Le recours au corpus permet néanmoins une finesse d'analyse de l'évolution, forme par forme, du système des démonstratifs, qui n'était pas envisageable auparavant. Il entraîne surprises, réévaluations et découvertes : *«* [...] le grand nombre des données qui nous sont désormais accessibles montre une situation fort inattendue en moyen français *»* (Marchello-Nizia, 1995, p. 165). Mais il en va de même pour l'opposition *that* / zéro, et pour la position des adjectifs en moyen anglais tardif.

C. Mair ajoute (1995, p. 260) : « [...] les innovations grammaticales généralement ne bouleversent pas le langage mais s'établissent d'abord dans des genres textuels spécifiques, des registres ou des niches fonctionnelles. Les corpus, comme témoignages de performance réelle, rendront plus faciles l'étude de ces types de contraintes. *»* Cette démarche est exemplifiée par l'étude de l'alternance *that* / zéro. Elle reste à entreprendre pour la position des adjectifs (seule la prose a été étudiée) et pour les démonstratifs. Il n'est pas exclu en effet que la distinction poésie / prose influence l'emploi des démonstratifs, en particulier pour la répartition entre déterminants et pronoms.

29.3 Acceptabilité et fréquence

Par définition, il n'existe pas, pour les états disparus d'une langue, de compétence du locuteur actuel. L'érudit contemporain ne saurait affirmer : cet énoncé n'est pas acceptable. En effet, sa connaissance de ce qui lui paraît possible ou non dans la période qu'il étudie provient uniquement de sa connaissance intime de textes en nombre fini, dont il a fini par abstraire les mécanismes lexicaux et syntaxiques dominants. Elle n'équivaut pas, loin s'en faut, à une capacité à produire des énoncés relevant de cet état de langue. La perception des régularités à l'œuvre est probablement

[202] D'où des précautions légitimes comme : « [...] après 1340, au moins en l'état actuel de notre documentation, on ne trouve plus aucune trace de ce morphème *cist* en français » (Marchello-Nizia, 1995, p. 159).

distordue, dans les deux sens : certains faits de très faible fréquence peuvent avoir échappé à l'attention et, à l'inverse, certaines caractéristiques dominantes peuvent être sous-estimées. L'oral est par ailleurs insaisissable, sinon par les biais qu'offrent certaines types d'écrits, avec le risque que rappelle C. Blanche-Benveniste (1997, p. 36) à propos de la *Grammaire des fautes* d'H. Frei de « confondre fautif et parlé », et de prendre « les fautes typiques de scripteurs inexpérimentés » pour des reflets de l'oral. La découverte de nouveaux documents, de nouvelles éditions critiques peuvent en plus amener à réévaluer la place de certains phénomènes[203].

Les corpus permettent par contre d'approcher les régularités centrales d'un état de langue oublié. Pour cerner les *«* impossibles de langue *»*, C. Marchello-Nizia (*ibid.*, p. 22) propose de recourir au raisonnement suivant : *«* On accordera [...] une importance privilégiée à l'absence de formes ou de constructions attendues, et corrélativement aux paraphrases. En effet, si un tour attendu n'est jamais attesté, et qu'on rencontre régulièrement sa paraphrase en lieu et place où on l'attendait, alors on a le droit de formuler l'hypothèse que le tour qu'on attendait là est, dans ce cas, agrammatical. *»*

La quantification occupe par conséquent une place centrale. Mais elle rencontre des difficultés sur des corpus d'états anciens de la langue. Lorsqu'il s'agit d'étudier des propriétés linguistiques *«* fines *»*, le nombre d'occurrences d'un phénomène donné dans une partie du corpus est souvent faible (inférieur à la dizaine). Il n'est d'ailleurs pas toujours possible, soit pour des raisons de coût soit plus fondamentalement parce que les sources sont lacunaires, de compléter les inventaires du phénomène visé. Ces petites quantités ne rendent cependant pas pour autant illégitime le recours à des modèles probabilistes appropriés pour évaluer leur significativité. Certains de ces modèles sont présentés au chapitre IX.

29.4 Affiner les explications

Le recours à des corpus diachroniques favorise pour l'analyse du système des démonstratifs en français un renouvellement de l'explication du changement morphologique. Traditionnellement, la causalité retenue était la suivante : un changement phonétique déclenche un changement morphologique qui peut lui-même entraîner un changement syntaxique. Les études récentes sur lesquelles s'appuie C. Marchello-Nizia poussent à relativiser dans ce cas le poids des changements proprement phonétiques (pour *ces*, par exemple). Parallèlement, les concordances facilitent l'étude détaillée des comportements syntaxiques (par exemple pour les formes préfixées en *i-*) et l'existence de textes enregistrés en nombre suffisant, une périodisation précise pour chaque forme (*cil* et *cist*

[203] « [...] les textes nous parviennent par copistes, et parfois générations de copistes interposées, auxquels s'ajoute inévitablement l'intervention de l'éditeur moderne ; jamais un texte n'est le pur reflet de l'usage de l'auteur ; il s'agit nécessairement d'une langue hybride [...] » (Marchello-Nizia, p. 22).

par exemple). Ces données et ces outils permettent de donner consistance aux facteurs qui sont invoqués : l'évolution de l'accent, qui passe du mot au groupe syntaxique, et l'influence de parentés de plus haut niveau, de systèmes méta-morphologiques et sémantiques généraux (avec la restructuration du système pronominal).

Nous avons vu l'usage de la notion d'analogie pour expliquer l'« invention » de *ce* : il viendrait compléter *ces* et faire pendant avec lui au couple *le* / *les*. C. Marchello-Nizia rappelle (*ibid.*, p. 176-178) les critiques qu'appelle l'usage de cette notion pour rendre compte, en dernière instance, de certaines évolutions[204]. L'analogie est le plus souvent utilisée au coup par coup. Elle fonctionne alors comme *«* explication » de la dernière chance. Elle est utilisée de manière *«* superficielle *»,* par opposition à des règles dûment formalisées.

Au delà des explications parfois hasardeuses par l'analogie, l'annotation linguistique de corpus étalés dans le temps fournit désormais la possibilité d'étudier des corrélations extrêmement complexes – et pratiquement non perceptibles sans appui informatique – entre des phénomènes situés aux différents niveaux de l'analyse linguistique ainsi que leur évolution au fil du temps. C'est le cas d'une des hypothèses majeures de C. Marchello-Nizia : la corrélation de l'évolution des démonstratifs avec celle des pronoms personnels. On souhaiterait alors tout naturellement dépasser le recours à des concordances et des comptages sur les seuls démonstratifs pour disposer de données chiffrées sur les deux systèmes et pouvoir examiner les corrélations, si elles existent, entre eux, par le recours, par exemple, à l'analyse multi-dimensionnelle (cf. chapitre IX). On progresserait vers le test effectif de l'hypothèse plus générale qui est posée (*ibid.*, p. 168) : *«* les systèmes morphologiques des langues s'organisent à un niveau supérieur en macro-systèmes sémantiques et formels plus abstraits, et ce sont ces méta-structures qui sont cause de certains des changements qui affectent les systèmes du niveau inférieur, immédiatement perceptibles, eux. *»* Dans une optique proche, les contraintes pesant sur l'omission du sujet pronominal en moyen français sont soumises dans (Dupuis *et al.*, 1992) à une analyse multivariable. À partir de l'examen de la distribution du sujet dans 10 textes s'échelonnant du premier tiers du XIVe siècle jusqu'à la fin du XV^{e} siècle, cette analyse montre que, parmi les facteurs examinés : la période du texte, l'opposition prose / poésie, le type de proposition et la personne du sujet, c'est le type de proposition dont l'influence ressort nettement : l'omission est plus souvent le fait des principales et des indépendantes que des enchâssées.

Les analogies réelles devraient être désormais plus facilement objectivables. La vision des causalités à l'œuvre dans le changement linguistique en sera probablement renouvelée. Ces causalités sont peut-être à chercher à des niveaux de structuration beaucoup plus abstraits (Kroch, 1990, p. 239) que ceux qui sont envisagés généralement.

[204] Cf. aussi (Kroch, 1990, p. 238)

CHAPITRE VI

D'UNE LANGUE A L'AUTRE : LES CORPUS ALIGNES

30. DEFINITION ET EXEMPLES

On appelle textes alignés (ou bi-textes) des couples de textes dont l'un est une traduction de l'autre et pour lesquels il existe un système de mise en relation entre segments du texte de « grain équivalent » : sections, paragraphes, phrases. On parle également de corpus bilingues.

Des occurrences de *guerre froide* ou *cold war* sont fournies par le Hansard aligné, c'est-à-dire les débats du Parlement canadien où la version en anglais est mise en correspondance avec la version française[205]. Voici quatre exemples de contextes alignés, où, à chaque fois, le texte source est anglais :

```
That is what is called leadership ,    | Voilà en quoi consiste le leadership .
not sticking one 's head in the sand   | Il faut éviter de faire l' autruche ,
, not looking through the rear - view  | de regarder en arrière et d' éprouver
mirror , not having some nostalgia     | une certaine nostalgie de l' ancienne
for the old cold war but saying it is  | guerre froide . Il faut plutôt se dire
time to make some change .             | que le moment est venu d'apporter des
                                       | changements.

This happened in 1990 , and now she    | C' était en 1990 . Aujourd'hui , elle
says : `` I do not understand why all  | dit qu' elle ne comprend pas pourquoi
of a sudden you are now saying we      | tout à coup nous trouvons à redire à
have a problem with the program '' ,   | ce programme . Mis à part le fait que
```

[205] Les contextes ont été fournis par L. Langlois (Dictionnaire canadien bilingue - Université d'Ottawa) utilisant sous licence TransSearch qui permet des concordances sur des textes alignés. TransSearch a été développé au CITI (Centre d'Innovations en Technologie de l'Information - Laval, Canada), devenu le RALI (Laboratoire de Recherche Appliquée en Linguistique Informatique). Cf. (Simard *et al.*, 1992).

quite apart from the fact that the geostrategic situation has changed tremendously in the period we are talking about . The cold war was pretty cold in 1990 .	la situation géostratégique a terriblement changé depuis , la guerre froide était plus que froide en 1990 .
I also want to acknowledge the staff reductions indicated by CSIS in the counterintelligence area . They are probably a function of the reduction in cold war intelligence battles that went on for many years .	Pour terminer , je voudrais parler de la réduction des effectifs mentionnée par le SCRS dans le secteur du contre - espionnage , réduction qui est peut - être attribuable à l' apaisement de la guerre froide .
It is not so easy to keep them in the cold dawn of post - war budgeting .	Il est moins facile de les tenir après la guerre , à l' époque froide des contrôles budgétaires .

On perçoit sur ces exemples, dont le second remotive les constituants de l'expression toute faite, les difficultés de la mise en correspondance (une phrase anglaise d'un côté, deux phrases françaises de l'autre dans l'exemple 2, l'inverse dans l'exemple 3). Le troisième exemple manifeste par exemple des décalages entre les deux versions (*intelligence battles that went for many years* est sans équivalent dans la version française). Le quatrième est une métaphore filée à partir de l'expression toute faite.

Ce bi-texte manifeste des types de contextes nouveaux par rapport à ceux examinés par Barkema (chapitre II) :

- *cold war* {nom}, où *cold war* est le modifieur du nom :

  cold war attack helicopters / hélicoptères d' assaut bons pour la Guerre froide

  cold war style helicopters / hélicoptères rappelant l'époque de la guerre froide

  cold war helicopter program / programme d' achat d' hélicoptères digne de la guerre froide

  the EH-101 cold war helicopters / hélicoptères EH-101 conçus pour la guerre froide

  cold war helicopters / hélicoptères de la guerre froide

  Ces contextes récurrents sont appuyés par la paraphrase suivante : helicopters to fight the cold war / hélicoptères destinés à la guerre froide ;

- des contextes qui précisent les parties prenantes du conflit larvé :

  the Moscow - Washington cold war / La guerre froide entre Moscou et Washington

  helicopters for the cold war with the Soviet Union / hélicoptères pour faire la guerre froide avec l' Union soviétique

  The cold war between the two blocs / cette guerre froide - là entre les deux Blocs

- *post cold war* {nom}, où le nom en question renvoie à une dimension temporelle, modifié par le syntagme *post cold war* :

  the post cold war environment / le climat d' après - guerre froide

  in a post - industrial , post - cold war world environment / en cette période postindustrielle et d' après - guerre froide

  In a post - industrial , post cold war environment / À l' ère postindustrielle , la guerre froide étant chose du passé

the post cold war era / dans l' ère de l' après - guerre froide

post cold war world / depuis la fin de la guerre froide

the post - cold - war situation / l' après - guerre froide

La version utilisée du Hansard aligné, qui correspond à trois ans de débats, représente 21,6 millions de mots anglais et 24,1 millions de mots français. Elle comprend 5 993 occurrences de *guerre*, 384 de *froide*, 5 977 de *war* et 673 de *cold*. Pour un volume globalement équivalent au corpus de Birmingham utilisé par Barkema, on rencontre près de trois fois plus d'occurrences de *cold war ou guerre froide* (314 occurrences). On ne trouve aucune occurrence de *guerres froides* ni de *cold wars*. On ne trouve qu'un seul exemple de discontinuité entre les deux composants de l'expression : c'est l'exemple 4 ci-dessus. Ces constats confirment l'analyse de Barkema sur la rigidité de l'expression. Dans 8 cas d'ailleurs, la traduction de *cold war* se fait par *Guerre froide*, la majuscule soulignant le fonctionnement comme un tout indécomposable.

31. UTILISATION DES TEXTES ALIGNES

Le recours aux textes alignés constitue par certains côtés une riposte aux limites rencontrées dans l'automatisation de la traduction automatique. Le point de départ n'est pas une formalisation de deux langues et de leur mise en correspondance, mais la réutilisation des traductions existantes produites par des traducteurs humains.

Les textes alignés fournissent un appui critique à la traduction. Cet appui peut consister à vérifier qu'il n'y a pas d'omissions dans la traduction. On en a précisément relevé une dans l'exemple 3 de la section 1. Un autre problème est celui des faux-amis partiels (Isabelle et Warwick-Amstrong, 1993, p. 302) : *Max fut arrêté par le FBI -> Max was arrested by the FBI* versus *Max arrêta le moteur -/-> Max arrested the engine, -> Max stopped the engine*. Disposer de contextes alignés permet de vérifier l'adéquation de la traduction qu'on se propose d'utiliser. Il importe alors de pouvoir filtrer les contextes sur des expressions des deux langues à la fois.

Les textes alignés servent de ressource pour les termes dont la traduction « homologuée » dans la langue-cible ne correspond pas forcément à une traduction mot à mot. Le Hansard aligné montre que les traducteurs utilisent généralement *droit compensateur* pour *countervail*, et parfois *droit compensatoire* (Isabelle, 1992). En langue générale, les textes alignés donnent accès à « la bonne expression » que le traducteur ne trouvera pas forcément dans un dictionnaire ou à des solutions auxquelles il n'avait pas pensé mais qui le satisfont et qui lui permettent de varier son expression. Voici quelques équivalences trouvées dans le Hansard pour l'expression *cartes sur table* (*ibid*.) :

Il a mis cartes sur table | He has put his facts on the table

Mettez-donc les cartes sur table | Put your cards on the table

Si c'est le cas, mettons cartes sur table [...] | If that is the case, let us get it on the table [...]

Peut-il jouer cartes sur table ? | Will he come clean with the Canadian people ?

Il devrait jouer cartes sur table avec les Canadiens | It should present Canadians with the straight goods.

Les techniques actuelles d'alignement poussent à vouloir exploiter le « trésor » que constituent les traductions déjà existantes. P. Isabelle (*ibid*.) indique : « Au Canada seulement, bon an mal an, le volume de traductions atteint au moins un demi-milliard de mots. [...] La masse des traductions produites chaque année contient infiniment plus de solutions à plus de problèmes que tous les outils de référence existants et imaginables. » L'objectif est alors de chercher s'il n'existe pas déjà une solution au problème de traduction rencontré, dans les traductions existantes, plutôt que d'en inventer une de toutes pièces. Les bi-concordanciers comme TransSearch permettent de telles recherches.

Les corpus alignés permettent de repérer des néologismes et la traduction qui en est donné. Ils viennent aussi remédier aux inévitables lacunes des dictionnaires. Gale et Church (1991) montrent par exemple que dans les corpus qu'ils avaient alignés, *en jeu* servait souvent de traduction à *at risk*, alors qu'un dictionnaire comme le *Robert et Collins* ne mentionne pas cette équivalence.

32. METHODES D'ALIGNEMENT

L'objectif est, selon P. Isabelle et S. Warwick-Amstrong (1993, p. 288) « la reconstitution automatique des correspondances traductionnelles qui unissent les segments d'un texte source et ceux de sa traduction. » Cet objectif est moins ambitieux que ceux qu'implique une traduction automatique : « Par opposition à la compétence active mise en jeu par les systèmes de traduction automatique, la recherche de correspondances dans les traductions préexistantes suppose seulement une compétence passive qui, en principe, devrait être moins difficile à atteindre » (*ibid.*, p. 289). La nature même de l'objectif conduit à des méthodes différentes. On part de « l'équivalence traductionnelle » qui est au contraire le résultat final escompté de la traduction automatique.

L'alignement peut s'effectuer aux différents niveaux de structuration de l'énoncé : des sections du texte aux mots en passant par les paragraphes et les phrases. C'est ce que P. Isabelle et S. Warwick-Amstrong (*ibid*.) nomment la « résolution » de l'alignement. Les correspondances deviennent de plus en plus difficiles à établir lorsqu'on diminue la taille des entités rapprochées. Les grandes sections d'un document sont général en relation bijective entre les deux versions. C'est encore souvent le cas pour les paragraphes. Les phrases font déjà exception. Une phrase dans une langue peut se traduire par deux phrases, voire plus dans l'autre, nous en

avons vu des exemples. L'ordre des propositions ou des phrases peut varier. En deça de la proposition, la variation de l'ordre des mots ainsi que le remplacement d'un mot dans une langue par une périphrase ou une expression polylexicale dans l'autre constituent des obstacles plus évidents encore à l'alignement.

P. Isabelle et S. Warwick-Amstrong (*ibid.*, p. 292) fournissent une définition tout à fait générale de l'alignement :

(T1, T2, Fs, C(Fs(T1), Fs(T2)))

T1 est le texte source, T2 sa traduction. Fs est une fonction de segmentation (cf. chapitres VII et VIII) qui fragmente le texte (il peut s'agit de mots, de phrases, de paragraphes, de sections). C est une fonction de correspondance qui relie l'ensemble des segments produits par Fs sur le texte source, Fs(T1), à l'ensemble des segments fournis par Fs sur le texte cible, Fs(T2).

Deux méthodes sont employées pour l'alignement. La première s'appuie sur l'existence d'une très forte corrélation entre la longueur d'un segment source et celle de sa traduction. La seconde utilise les paires particulières des mots pour mettre en corrélation. D'autres propositions sont des variations sur ces propositions de base ou encore la combinaison des deux approches.

La première méthode utilise donc la « corrélation très forte entre la longueur des segments qui sont mis en correspondance traductionnelle » (*ibid.*, p. 295). Les segments peuvent être mesurés en nombre de mots (Brown et al., 1991) ou en nombre de caractères (Gale et Church, 1991)[206]. Chacun des deux textes est d'abord décomposé en phrases[207]. On se donne un ensemble d'appariements licites (un / zéro, zéro / un, un / un, un / deux, deux / un, etc.). Dans la plupart des cas, on n'autorise pas les liens croisés. On examine alors tous les appariements possibles compatibles avec les appariements retenus comme licites. On calcule un score reflétant la qualité des corrélations des longueurs des segments contenus pour chaque appariement. On retient l'appariement dont le score est le meilleur. Les résultats sont entre 95 et 100 % d'appariements justes. Cette famille de méthodes présente l'avantage de ne pas nécessiter de recours à un dictionnaire. Inversement, l'examen « à gros grain » des corrélations entre les deux textes empêche une resynchronisation quand l'appariement se décale à un endroit donné.

La deuxième méthode prend appui sur les mots apparentés entre deux langues proches (*gouvernement* / *government* par exemple). Il ne s'agit pas d'utiliser un dictionnaire mais de repérer des distances entre chaînes de caractères (par exemple en termes de coût de passage d'une chaîne à l'autre en nombre d'effacements, ajouts et substitutions).

[206] Cf. aussi (Blank, 1995 ; Langé et Gaussier, 1995).
[207] Tâche qui est moins évidente qu'elle n'en a l'air. Que l'on pense aux titres, aux énumérations, aux légendes de tableaux et de figures, aux incises.

33. PROBLEMES ET ENJEUX

P. Isabelle et S. Warwick-Amstrong insistent (*ibid.*, p. 290) sur la « compositionnalité de la traduction » : « la traduction d'une unité textuelle est généralement fonction de la traduction des parties de cette unité, et ce, jusqu'au niveau d'un ensemble fini d'équivalences élémentaires. » C'est effectivement ce principe qui rend possible la démarche d'alignement. Mais en même temps, comme nous l'avons vu, la « résolution » de l'alignement peut être plus ou moins grande : des correspondances des grandes parties du texte et des paragraphes s'accommodent de décalages à un niveau plus fin (c'est le cas du troisième exemple de la section 1, où une partie de la phrase source n'a pas de correspondant traductionnel). Comme l'indiquent P. Isabelle et S. Warwick-Amstrong (*ibid.*, p. 302), un système d'alignement fin permettrait de repérer les erreurs de traduction liés aux faux amis, c'est-à-dire les cas où un mot est traduit par un mot trompeusement proche (comme *eventually* pour *éventuellement*).

Les textes alignés permettent également d'examiner les équivalences entre séquences non compositionnelles : les décalages localisés qu'elles représentent sont contrebalancés par l'alignement des structures plus vastes dans lesquelles elles figurent. Les textes alignés permettent en ce sens une répartition relativement harmonieuse des tâches entre « machine » et traducteur. L'alignement produit un dégrossissage des mises en correspondance. En fonction de la requête qu'il effectue, le traducteur puise dans les réponses et s'appuie sur les blocs alignés pour examiner les parallèles ou les divergences dans le détail. L'alignement produit automatiquement est évidemment limité, mais il est suffisant pour beaucoup de tâches de traductique.

L'alignement, du moins « à gros grains »[208], peut sembler une tâche plus aisée que l'étiquetage ou le parsage. En tout cas, il y a un grand décalage entre la relative simplicité des méthodes employées pour obtenir des textes alignés et la richesse extrême des utilisations de ces corpus bilingues. Ce décalage même est source d'espoir.

[208] Par opposition à un alignement syntagme à syntagme voire mot à mot.

TROISIEME PARTIE

METHODES ET TECHNIQUES

CHAPITRE VII

CONSTITUER UN CORPUS

34. DEFINITIONS ET TYPOLOGIE DES CORPUS

Il y a vingt ou trente ans, la constitution d'un corpus électronique était une tâche ardue : saisie et correction du texte sur cartes perforées, traitement informatique dans des centres de calcul distants, sur des machines dont les capacités de stockage et de calcul limitaient la taille des données manipulables ... Avec l'avènement de la micro-informatique, l'introduction des réseaux, l'augmentation de la taille des mémoires et la rapidité croissante des traitements, la situation a radicalement changé. Beaucoup d'écrits professionnels existent directement sous forme électronique et sont donc « recyclables » au sein d'un corpus. Le « captage » de textes est désormais aisé.

Paradoxalement, la notion même de corpus s'en est obscurcie. À l'orée des traitements informatiques de données textuelles, le coût même de la création d'un corpus conduisait à peser mûrement les textes à y intégrer, à identifier précisément les critères de rassemblement. Aujourd'hui que le texte électronique foisonne, des documents se trouvent parfois agrégés avant tout parce qu'ils sont faciles d'accès[209], sans que leur mise en relation ait été réellement pensée. La mûre pesée d'un regroupement adéquat à l'objectif poursuivi cède le pas à la seule disponibilité des ressources. La communauté

[209] Ce qui est appelé crûment dans (Marcus *et al.*, 1993, p. 313, n. 1) des regroupements « opportunistes ».

du TALN appelle souvent corpus les grandes collections de documents qui lui servent à mettre au point ses traitements. Les rencontres organisées depuis plusieurs années par l'ACL (Association for Computational Linguistics) sur les « très grands corpus » (*very large corpora*) traitent de très vastes données textuelles plutôt que de corpus à proprement parler. On serait plutôt tenté de voir là « du texte », texte dont on ne sait pas toujours très bien de quels usages langagiers il est représentatif.

Nous adoptons la définition plus restreinte de John Sinclair (1996, p. 4) : « Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage. » C'est à dessein que le mot « texte » n'est pas employé ici. En effet, comme pour ***Archer*** ou pour ***BNC***, les techniques d'échantillonnage peuvent amener à briser la séquentialité des textes de départ : on extrait éventuellement des fragments en plusieurs endroits d'un même texte pour éviter de sur-représenter ou sous-représenter certaines caractéristiques[210]. Les *corpus de textes* (complets) s'opposent aux *corpus d'échantillons* (*ibid*., p. 9). On cherche en outre à respecter les critères suivants : une taille aussi importante que les moyens techniques le permettent[211] (par souci de représentativité), des échantillons diversifiés (et éventuellement de taille similaire), une origine nettement repérée (les coordonnées des documents primaires sont conservées). Par opposition (*ibid.*) , « [d]es mots comme 'collection' ou 'archive' renvoient à des ensembles de textes qui ne nécessitent pas de sélection ou d'organisation, ou dont la sélection ou l'organisation ne nécessitent pas de critères linguistiques[212]. » Les CD-ROM du journal *Le Monde*, par exemple, rassemblent des articles relevant de discours parfois éloignés (langue générale de la vie politique et sociale – nationale et internationale, langues spécialisées diverses : économie, sport, météorologie, etc.). Il est donc plus adéquat de parler de « la collection du *Monde* sur CD-ROM » que du « corpus du *Monde* ».

On peut alors opposer *corpus de référence* et *corpus spécialisé* : « Un corpus de référence est conçu pour fournir une information en profondeur sur une langue. Il vise à être suffisamment étendu pour représenter toutes les variétés pertinentes du langage et son vocabulaire caractéristique, de manière à pouvoir servir de base à des grammaires, des dictionnaires et d'autres usuels fiables » (*ibid.,* p. 10). ***Brown***, ***LOB*** et ***BNC*** constituent des corpus de référence, les deux premiers uniquement pour l'écrit, le troisième

[210] Par exemple, les phrases analysées manuellement à l'université de Lancaster (1 million de mots) dans le cadre de la collaboration avec IBM Watson (Black *et al.*, 1993, p. 23) ont été extraites au hasard d'un ensemble de 20 millions de mots de dépêches de l'agence Associated Press. Elles ne sont pas consécutives, ce qui ne facilite d'ailleurs pas forcément leur compréhension par les annotateurs.

[211] John Sinclair ajoute : « Un corpus est supposé contenir un grand nombre de mots. L'objectif fondamental de la constitution d'un corpus est le rassemblement de données en grandes quantités ». Il se garde de préciser ce qu'il entend par grandes quantités ...

[212] G. Leech fait écho (1991, p. 11) : « [...] en fin de compte, la différence entre une archive et un corpus doit résider dans le fait que ce dernier est conçu ou nécessité pour une fonction 'représentative' précise. »

pour l'oral également. Les deux premiers ne répondent d'ailleurs plus aux exigences de taille qui peuvent être les nôtres aujourd'hui. Les *corpus comparables* (*ibid.*, p. 12) constituent des sélections de textes similaires dans plus d'un langage ou dans plusieurs variétés d'un langage. On peut considérer ***LOB*** et ***Brown*** comme des corpus comparables. Tous deux regroupent des textes provenant des mêmes « genres » et de la même année : 1961, mais ils relèvent pour le premier de l'anglais, pour le second de l'américain. Les *corpus spécialisés* sont limités à une situation de communication, ou à un domaine. Parmi ces corpus, on trouve les ensembles relevant de sous-langages que l'on trouve dans les domaines scientifiques et techniques (cf. section 3). Les *corpus* ou *collections parallèles* sont constitués d'un ou de plusieurs documents traduit(s) dans une ou plusieurs langues (cf. chapitre VI). L'exemple canonique est le *Hansard* : les débats du Parlement canadien, en anglais et en français.

Beaucoup de corpus constituent des ressources achevées, dès lors immuables : on n'y ajoute plus rien, mais on peut en extraire éventuellement des sous-corpus (l'oral dans ***BNC*** par exemple, ou une diachronie restreinte dans ***Archer***). À l'inverse, avec la possibilité de « capter » en continu des données dans certains secteurs (les fichiers de composition de grands journaux comme le *Times*, par exemple), est apparue la notion de *corpus de suivi*[213] – *monitor corpus* (Sinclair, 1996, p. 4). Par définition, un tel corpus ne cesse de croître. Il devient alors possible d'étudier l'évolution de certains phénomènes langagiers : néologismes, emplois privilégiés à un moment donné de certains suffixes ou préfixes, etc., un peu comme les éditions papier de certains dictionnaires d'usage (*Le Petit Larousse*, *Le Petit Robert*) servent de « sonde » sur le lexique et ses changements. Dans la mesure où ces corpus de suivi sont récents, ils ne peuvent renseigner dans l'immédiat que sur la courte durée (moins d'une décennie). Mais avec le temps, ils contribueront à notre connaissance de l'évolution de certains secteurs de la langue (cf. chapitre V).

« Un corpus électronique est un corpus qui est encodé de manière standardisée et homogène pour permettre des extractions non limitées à l'avance » (*ibid.,* p. 5). En effet, la simple existence sur support électronique ne fait pas d'un ensemble de textes un corpus électronique. Encore faut-il que ce document obéisse à des conventions de représentation, de codage répandues, voire faisant consensus, qui permettent la transmission et la réutilisation des données textuelles en cause (cf. section 5).

[213] ou encore *corpus baromètre*.

35. LANGUE GENERALE

35.1 Etudier une dimension particulière

La nature des phénomènes à étudier peut réclamer des données très vastes ou au contraire se satisfaire d'un corpus restreint. H. Barkema (1994, p. 271) indique ainsi : « [...] un corpus d'un million de mots est bien trop restreint pour étudier la flexibilité [des expressions toutes faites] et [...] un corpus de 20 millions de mots est trop petit pour trouver un nombre suffisant d'occurrences de toutes les expressions [idiomatiques]. » Il fournit les chiffres suivants (1993, p. 271-272) : sur l'ensemble des noms composés répertoriés par *LDOCE* (*Longman Dictionary of Contemporary English*), 88 % d'entre eux apparaissent une fois ou plus dans les 20 millions de mots du corpus de Birmingham, 48 % plus de 10 fois et 30 % plus de 20 fois. La proportion de ceux d'entre eux pour lesquels une étude de flexibilité est possible s'avère donc réduite.

Donnons un exemple de corpus spécialisé, conçu pour l'étude d'un phénomène bien délimité. G. Engwall (1994, p. 60-64) se fixe comme objectif, au milieu des années soixante-dix, d'étudier sur le plan linguistique les mots, les syntagmes et les constructions de la prose française littéraire contemporaine, à travers le roman. Après avoir considéré l'état des ressources électroniques de l'époque (et en particulier le corpus du *Trésor de la Langue Française*), G. Engwall retient la période 1962-1970, pour pouvoir rendre compte des années soixante. La dénomination de « roman » recouvrant des écrits bien divers, le classement d'une bibliographie française, *les Livres de l'année*, lui sert de pierre de touche. Les listes des meilleures ventes des *Nouvelles littéraires* et du *Figaro littéraire* constituent un filtre supplémentaire. Environ 400 titres répondent à ces premiers critères de période, de genre et de diffusion. L'élimination des livres traduits ou de ceux dont la première édition précède le début de la période retenue ramène cet ensemble à 161 titres. Deux conditions supplémentaires sont retenues : l'auteur doit être né en France et faire partie des auteurs les plus jeunes des meilleurs ventes, l'action du roman doit être située dans la France de l'après-guerre (ce qui nécessitait un examen des textes). Dernière contrainte : la taille globale du corpus, fixée à 500 000 mots (par comparaison avec des recherches similaires). D'où le choix de fragments totalisant 20 000 mots (la taille d'un livre de poche très court) pour chacun des 35 romans finalement choisis. Pour mieux rendre compte de chacune des œuvres, ces fragments ne sont pas consécutifs : ils sont formés de 10 échantillons de 2 000 mots extraits au hasard de chacune des œuvres.

35.2 Constituer un corpus de référence

Deux positions s'opposent et constituent les pôles entre lesquels se répartissent les créateurs de corpus. « Gros, c'est beau » (*more data is better data*), pourrait être le slogan de la première. La conviction sous-jacente est que l'élargissement mécanique des données mémorisables (les centaines de millions de mots actuelles deviendront à terme des milliards) en fait inévitablement un échantillon de plus en plus représentatif du langage traité. Si l'on n'arrive pas à cerner précisément les caractéristiques de l'ensemble des productions langagières, il ne reste qu'à englober le maximum d'énoncés possibles. À terme, la nécessité de choisir finirait par s'estomper.

La seconde approche, plus sensible aux variations propres aux données textuelles, constitue des ensembles aux conditions de production et de réception plus nettement définies et corrélées à leurs caractéristiques langagières. La logique de cette position conduit même à « équilibrer » en taille les échantillons retenus, voire à ne pas retenir des empans de texte continus, de manière à éviter de sur-représenter des « lieux » du texte particuliers (l'introduction par exemple). Cette technique de constitution des textes par échantillonnage est souvent pratiquée pour les corpus anglo-saxons (***BNC***, ***Archer***, ***LOB***, ***Brown***, ***Helsinki***). L'échantillonnage touche donc à la fois le choix des documents à intégrer et la partie de ces documents à conserver. Biber (1993a, p. 222-226) montre les variations des pondérations de certains traits linguistiques selon le genre considéré. Les fréquences des étiquettes possibles pour un mot changent. Dans ***LOB***, pour les textes de fiction, *known* est un passif dans 26 % des cas, un prétérit dans 65 %, et un adjectif dans 6 %. Ces proportions passent à 65 %, 13 % et 15 % respectivement pour les textes « expositifs » (*exposition*). Les prédictions que l'on peut faire sur la catégorie la plus probable pour *known* dépendent donc du genre choisi pour estimer les fréquences des catégories possibles[214]. Il en va de même pour la probabilité d'une catégorie lorsqu'on connait la catégorie précédente. Dans le même corpus, la copule *be* est suivie d'un passif dans 13 % des cas dans les textes de fiction et dans 31 % des cas dans les textes « expositifs ». Biber et Finegan (1994), sur un corpus d'articles *du New England Journal of Medicine* et de *The Scottish Medical Journal*, montrent également que les parties canoniques d'un article scientifique (introduction, méthodes, résultats, discussion) comportent des différences sensibles entre elles. Le présent est fréquent dans l'introduction et la discussion et relativement rare dans la partie méthodes. Le passé a la distribution inverse. On comprend dès lors mieux la politique qui consiste à « démembrer » certains documents pour ne pas sur-représenter certaines de leurs sous-parties, et plus largement cette « échantillonnite » qui surprend souvent un esprit français.

[214] A. Voutilainen dans (Karlson *et al.*, 1995), montre que, dans les corpus « équilibrés » entre différents genres que sont ***Brown*** et ***LOB***, *cover* (*couvrir*, *couverture*) est un nom dans 40 % des cas, un verbe dans 60 %. Dans un manuel d'entretien de voiture, il s'agit dans tous les cas d'un nom.

La démarche suivie pour la constitution de ***BNC*** (Burnard, 1995), conçu pour être un corpus de référence pour l'anglais, s'inscrit totalement dans cette seconde optique, à ceci près que les registres ne sont pas pris en compte. Les critères de choix diffèrent pour l'écrit et pour l'oral. En ce qui concerne l'écrit, plusieurs contraintes se superposent :

- le domaine : 75 % de textes « informatifs », le reste appartenant à la fiction ;
- le support : 60 % de livres[215], 30 % de périodiques, le reste comprenant des écrits non publiés ou des supports de discours (écrits pour être lus, comme les informations radio-télévisées) ;
- la datation : les ouvrages de fiction de 1960 à 1993 (pour tenir compte de leur durée de vie plus grande) et les ouvrages « informatifs » de 1975 à 1993 ;
- la diffusion : une liste de livres imprimés disponibles, les listes des meilleures ventes, celles de prix littéraires, les indications de prêts en bibliothèque (à la fois les ouvrages les plus prêtés et les ouvrages en prêt à court terme, qui sont donc très demandés) ont ainsi servi à choisir des livres « bien diffusés ».

Pour l'oral, l'objectif est la conversation spontanée. Le corpus est constitué par échantillonnage démographique en termes d'âge, de sexe, de groupe social et de région. Les 124 personnes choisies sur ces critères et à partir d'un entretien, âgées d'au moins 15 ans, disposaient pendant quelques jours d'un magnétophone portable pour pouvoir enregistrer leurs conversations. Les consignes étaient de varier les moments d'enregistrement (jours ouvrés / fins de semaine) et de noter à chaque fois la situation d'interlocution (datation, environnement, participants). L'enregistrement pouvait être effectué à l'insu des participants par la personne choisie, mais les interlocuteurs étaient prévenus *in fine* pour que l'on puisse effacer l'enregistrement si l'anonymat réalisé ne leur suffisait pas. En tout, plus de 700 heures d'enregistrement ont été réalisées. Outre cet échantillon démographique, ont été intégrées des transcriptions d'interactions orales typiques dans divers domaines : affaires (réunions, prises de parole syndicales, consultations médicales ou légales), éducation et information (cours et conférences, informations radio-télévisées), prises de parole publiques (sermons, discours politiques, discours parlementaires et légaux), loisirs (commentaires sportifs, réunions de clubs).

35.3 Peut-on constituer des échantillons représentatifs ?

Les deux positions exposées en 2.3 s'accordent implicitement sur la difficulté, en matière de langage, à donner une définition positive de la

[215] Les extraits de livres représentent 45 000 mots d'un seul tenant, le début étant choisi au hasard (en respectant toutefois les limites discursives du type chapitre).

représentativité[216]. Veut-on représenter les textes effectivement reçus ? Ou bien les textes et autres énoncés produits ? Les genres et domaines fournissent pour l'écrit un découpage, insatisfaisant certes, mais utilisable, des types à représenter. Pour l'oral, l'identification des classes à considérer est moins avancée. Notre connaissance de la « population » des données langagières est donc encore extrêmement fragmentaire. Les erreurs statistiques classiques sont par conséquent monnaie courante : l'échantillon est trop petit pour bien représenter la population, l'échantillon est systématiquement biaisé – il s'écarte significativement des caractéristiques de la population (Biber, 1993a, p. 219-220).

36. LANGUES DE SPECIALITE ET SOUS-LANGAGES

À l'opposé de la langue générale que cherchent à représenter les corpus de référence, se trouvent les usages spécialisés. Les dénominations (langues spécialisées, langues de spécialité, sous-langages) impliquent des analyses et des visées différentes. Parler de langue spécialisée, n'est-ce pas insister sur la continuité entre la langue générale et ce fonctionnement particulier ? La notion de langue de spécialité met plutôt l'accent sur le domaine technique ou scientifique concerné. Par sous-langage, Harris entend un fonctionnement langagier tout à fait spécifique.

36.1 Les hypothèses de Z. Harris

Z. Harris, à partir du milieu des années soixante-dix et jusqu'aux années quatre-vingt dix, oppose le caractère relativement flou des restrictions qu'un opérateur donné impose à ses arguments en langue générale (l'argument de *mourir* peut être un nom +animé, mais aussi un nom abstrait : *la mort d'une illusion*) aux limites extrêmement nettes rencontrées[217] dans ce qu'il appelle les sous-langages[218] : langages de disciplines scientifiques ou techniques, méta-langage (comme celui de la grammaire ou de la linguistique). Selon lui, ces sous-langages se caractérisent par un lexique limité et par l'existence de schémas de phrases en nombre fini. Ces schémas ont la particularité d'être

[216] On se reportera à (Biber, 1993a, 1994) pour une discussion approfondie.
[217] « Le caractère distinctif d'un sous-langage, c'est que pour certains sous-ensembles des phrases du langage, les restrictions de sélection, pour lesquelles on ne peut pas fournir de règles pour le langage dans son ensemble, intègrent la grammaire. Dans un sous-langage, les classes lexicales ont des frontières relativement tranchées qui reflètent la division des objets du monde en catégories qui sont clairement différenciées dans le domaine » (Sager, 1986, p. 3).
[218] (Harris *et al.*, 1989) fournit à la fois le cadre méthodologique global et des exemples d'analyses effectives, en particulier sur le français (elles sont dues alors à A. Daladier).

des combinaisons particulières de sous-classes de mots propres au sous-langage en question. Ainsi, dans ***Menelas***, sous diverses formulations se manifeste le schéma N1 dilater N2, où N1[219] ressortit à la classe des médecins et N2 à celle des artères : on dilate une artère coronaire, une artère circonflexe, etc[220].

La dénomination *sous-langage* tient du faux-ami. Ces sous-langages ne sont pas forcément en effet des sous-ensembles de la langue générale. Certains traits de la langue générale s'y retrouvent, d'autres leur sont propres. La prédictibilité de certains arguments peut provoquer leur omission systématique (on ne parlera pas ici d'ellipse) : par exemple, dans le domaine de la vinification, *on sucre* est acceptable, mais **on sucre le moût*, qui explicite l'argument, n'est pas un énoncé bien formé. Inversement, les sous-langages peuvent recourir à des patrons syntaxiques particuliers qu'il serait difficile d'intégrer tels quels à une grammaire « de langue »[221]. C'est le cas de certains motifs dénominatifs qui forment de véritables « grammaires locales ». Par ailleurs, les sous-langages diffèrent des langages contrôlés. Ils résultent d'ajustements lents et pour une large part non raisonnés au sein d'une communauté langagière restreinte. Les langages contrôlés se caractérisent également par un lexique et une syntaxe limités, mais ils proviennent d'une « planification » linguistique dans des domaines où une communication moins équivoque ou plus concise est particulièrement importante (dans l'aviation, par exemple).

36.2 Analyses de sous-langages

36.2.1 La méthodologie harrissienne

Cette vision des sous-langages s'accompagne d'une méthode pour mettre au jour les classes de mots et les patrons syntaxiques caractéristiques d'un sous-langage. Pour reprendre les termes de N. Sager (1987, p. 198) : « Si l'on applique à un corpus de textes d'un secteur scientifique des méthodes de linguistique descriptive similaires à celles utilisées pour le développement d'une grammaire d'une langue dans son ensemble, on obtient des motifs précis de cooccurrences de mots à partir desquels on peut définir des sous-classes de mots et des séquences de ces sous-classes qui sont caractéristiques (c'est-à-dire une grammaire). Ces catégories lexicales et formules syntaxiques de la grammaire du sous-langage sont étroitement corrélées aux classes d'objets du monde et aux relations qui sont propres à

[219] N1 n'est pas toujours exprimé, par exemple dans la nominalisation *dilatation de N2* ou dans l'utilisation du passif *N2 a été dilaté*.
[220] Il s'agit d'ailleurs d'une métonymie, c'est en fait un segment qui est dilaté et non l'artère entière.
[221] Les manuels informatiques anglais ont par exemple un emploi particulier de *to vary on [un dispositif]*, signifiant approximativement *le mettre en marche* dans des phrases comme « The system will be unable to vary on the device » (Black *et al.*, 1993, p. 112).

ce sous-domaine. Ils fournissent donc un ensemble de structures sémantiques pour refléter les connaissances de ce domaine. » L'objectif est ainsi résumé (*ibid.*, p. 198) : « La grammaire d'un sous-langage doit 'attraper' les restrictions d'occurrences qui distinguent un champ de discours scientifique d'un autre. »

Les étapes de cette mise en évidence sont les suivantes. En premier lieu, une analyse syntaxique (manuelle pour Harris, automatique pour Sager) d'un corpus du sous-langage considéré. En second lieu, une régularisation syntaxique par mise en phrases élémentaires (de type sujet – verbe – compléments éventuels). Cela suppose des restructurations et transformations linguistiquement fondées (passage d'une nominalisation au verbe correspondant : *dilatation d'une artère coronaire* / *X dilate une artère coronaire*, passage à l'actif pour les passifs, etc.) de manière à augmenter les proximités. L'interrogation d'un expert du domaine[222] permet de disposer des entités (arguments de verbes) qui lui paraissent fondamentales. Sur cette base, les régularités opérateur / arguments (verbe / sujet et compléments) permettent de mettre au jour les classes et les schémas caractéristiques du sous-langage.

36.2.2 Les analyses réalisées dans ce cadre

Les travaux fondateurs sont ceux de Harris et de son équipe sur le discours pharmaceutique et biologique (Harris *et al.*, 1989 ; Ryckman, 1990) ainsi que ceux de l'équipe de N. Sager (New York University), sur le langage médical (Sager *et al.*, 1987), ces derniers s'appuyant sur un parseur de l'anglais. L'examen d'autres domaines est rapporté dans (Grishman et Kittredge, 1986). La communauté du TALN, tant anglo-saxonne que française, s'est souvent inspirée de l'approche harrissienne des sous-langages pour traiter les domaines restreints auxquels elle est souvent confrontée.

36.3 Evaluation et perspectives

Curieusement, en France, dans la communauté linguistique, la conception harrissienne des sous-langages a eu peu de postérité, en dehors des travaux d'Anne Daladier (1990). Les travaux autour de Maurice Gross, disciple de Harris, se sont centrés sur les propriétés des entrées lexicales de la langue

[222] Cf. (Daladier, 1990, p. 75) : « Les catégories d'analyse du contenu informatif de ces textes ont été pour la plupart induites, en employant des méthodes d'analyse distributionnelles, de la formulation de l'information dans ce domaine. Seules les catégories 'élémentaires', c'est-à-dire celles dont le sens ne dépend pas d'autres catégories, et qui sont représentées pour cette raison comme des arguments terminaux de catégories ou de combinaison de catégories de niveau supérieur, ont été directement introduites par des experts du domaine (*i.e.* de façon non constructive. » D'autres travaux menés dans cette optique se sont inspirés de nomenclatures existantes en médecin.

générale. En outre, l'accent porte sur une caractérisation avant tout syntaxique : la sémantique est conçue comme trop peu formalisable[223], alors que les travaux de Harris sur les sous-langages aboutissent à des « grammaires sémantiques » qui associent aux différentes positions de patrons syntaxiques des classes sémantiques restreintes. L'Analyse Automatique du Discours (AAD), développée par Michel Pêcheux (Pêcheux, 1969 ; Maingueneau, 1991) au début des années soixante-dix a utilisé une méthode de normalisation manuelle des énoncés, elle aussi inspirée de l'analyse distributionnelle, et assortie d'un traitement informatique. L'accent était mis cependant sur la langue générale, ou du moins sur des domaines non techniques (discours politique). Les recherches contemporaines sur les sous-langages ne sont pas citées.

Aujourd'hui, comme le chapitre II l'a montré, l'existence d'analyseurs robustes rend partiellement possible l'application à grande échelle de la méthodologie harrissienne. On peut attacher automatiquement à de vastes documents des arbres syntaxiques, y compris en utilisant des méthodes d'apprentissage pour adapter le parseur à certains phénomènes propres aux documents en cause (sous-catégorisation des adjectifs, attachements prépositionnels). Les arbres syntaxiques peuvent être simplifiés pour obtenir des phrases élémentaires. Des opérations de réécriture d'arbres peuvent, en fonction du matériel lexical de l'arbre, transformer encore ces arbres (passage du passif à l'actif etc.) pour faciliter la mise en évidence de régularités. Ce nouveau contexte permet surtout d'examiner trois questions.

Tout d'abord, les énoncés d'un domaine particulier, qui relèvent donc pour Harris d'un sous-langage, présentent-ils vraiment des particularités syntaxiques par rapport à la langue dite générale, à la fois en ce qui concerne les constructions rencontrées et les types de contraintes syntaxiques des entrées lexicales ? L'existence de vastes corpus de référence, au sens donné en section 1, autorise des études contrastives nouvelles sur ce point.

En second lieu, Harris s'appuyait sur un informateur du domaine et utilisait les catégories d'entités fournies par cet informateur comme point de départ pour déterminer les classes d'opérandes en fonction des opérateurs utilisés. Cependant, une partie des recherches actuelles en TALN qui visent à dégager, à partir d'une analyse syntaxique, les opérateurs et leurs arguments au sein d'un domaine donné, essaient souvent de le faire sans ce recours à un premier dégrossissage conceptuel du domaine. L'économie de ce recours s'explique en partie par la difficulté d'obtenir ce type de renseignements : on dispose parfois de textes d'un domaine spécialisé, mais pas forcément d'informateurs compétents dans ce domaine. Existe aussi la conviction qu'il

[223] Les travaux plus récents autour de Gaston Gross sur les « classes d'objets » (Gross, 1994 ; Le Pesant, 1994) nous semblent également éloignés de l'optique ouverte par l'hypothèse des sous-langages. Il s'agit de catégoriser les mots en fonction des classes d'opérateurs qui leur conviennent : ainsi un bruit sera plutôt un événement que quelque chose de concret dans la mesure où l'on dit : « un bruit se produisit », Malgré cet emploi de la notion harrissienne d'« opérateur approprié », deux divergences essentielles demeurent : l'hypothèse que l'on peut isoler de telles classes en langue générale ; le recours à l'intuition du linguiste et non à un corpus.

suffit de disposer d'un ensemble suffisamment vaste de documents du domaine pour que le retraitement d'analyses syntaxiques fasse émerger les régularités syntactico-sémantiques. La question demeure donc : peut-on induire les schémas d'un domaine sans le recours à une expertise humaine, soit au départ, soit pour valider les regroupements produits automatiquement ? Bouaud *et al.* (1997), pour ***Menelas***, comparent les résultats des classements inspirés de la méthodologie harrissienne avec une nomenclature médicale « à gros grain ». Ils aboutissent à un constat nuancé : les regroupements sur la base de contextes syntaxiques élémentaires sont relativement proches des classes de cette nomenclature, mais il est nécessaire de faire appel à des connaissances du domaine pour préciser ou corriger cette catégorisation à base linguistique.

En troisième lieu, les travaux sur les sous-langages traitent souvent tous les discours produits dans un domaine comme utilisables au même degré par la méthode d'analyse proposée. Dans le domaine médical, par exemple, on trouve cependant différents types de textes, qui correspondent à des situations de communication typiques : manuels (destinés au futur médecin), compte-rendus d'examens ou de traitements, lettres à des collègues sur un patient commun, mais aussi articles scientifiques sur de nouveaux traitements, vulgarisation, etc. Les trois premiers types seuls se trouvent représentés dans ***Menelas***. L'analyse séparée de ces trois types montre que le discours didactique n'est pas forcément, au moins dans ce cas, le meilleur « observatoire » des régularités de ce domaine : par souci de généralisation, il utilise des hyperonymes qui ne se rencontrent pas dans les compte-rendus d'hospitalisation. On y trouve peut-être des régularités propres à tout discours didactique (pluriels génériques, présent de vérité générale, etc.) qui « parasitent » la perception du sous-langage proprement dit. Dernière question donc : comment articuler finement sous-langages et genres discursifs ?

37. ARTICULER TYPOLOGIE INTERNE ET TYPOLOGIE EXTERNE

La méthodologie à suivre pour délimiter l'ensemble que l'on souhaite représenter et pour rassembler des matériaux effectivement représentatifs combine, pour le moment encore très empiriquement, une caractérisation des situations de communication pertinentes, des genres et registres utilisés et des types de textes en circulation.

37.1 Typologie des textes, genres et registres

D. Biber distingue clairement les types de textes, qui relèvent de l'analyse linguistique, et les registres ou « genres », qui correspondent à une catégorisation sociale. Pour lui, les types de textes correspondent à des corrélations de caractéristiques linguistiques qui participent d'une même fonction globale. Ils ne se confondent ni avec les typologies fonctionnelles ni avec les « genres ». Les genres ou registres sont les catégories intuitives qu'utilisent les locuteurs pour répartir les productions langagières. On l'a vu à propos de ***Brown*** ou d'***Archer***, elles mêlent un repérage thématique à gros grain (Médecine, Science) et une utilisation de « formes de textes » (théâtre, sermons et homélies, journaux intimes). Ces catégories évoluent au fil du temps. Elles fournissent néanmoins un premier découpage des catégories de textes à prendre en compte.

37.2 Typologie des paramètres situationnels

D. Biber (1994, p. 380-385) fournit un certain nombre de paramètres situationnels permettant de décrire les documents intégrés dans un corpus :

1. Canal : écrit / parlé / écrit lu
2. Format : publié / non publié
3. Cadre : institutionnel / autre cadre public / privé-interpersonnel
4. Destinataire :
    a. pluralité : non compté / pluriel / individuel / soi-même
    b. présence : présent / absent
    c. interaction : aucune / peu / beaucoup
    d. connaissances partagées : générales / spécialisées / personnelles
5. Destinateur :
    a. variation démographique : sexe, âge, profession etc.
    b. statut : individu / institution dont l'identité est connue
6. Factualité : informatif-factuel / intermédiaire / imaginaire
7. Objectifs : persuader, amuser, édifier, informer, expliquer, donner des consignes, raconter, décrire, enregistrer, se révéler, améliorer les relations interpersonnelles, ...
8. Thèmes : ...

Attacher les valeurs de ces paramètres au corpus constitué permet d'examiner le lien entre cet ancrage situationnel et la caractérisation proprement linguistique du corpus.

38. NORMALISER UN CORPUS

L'échange des corpus et leur réutilisation ont buté jusque récemment sur l'éclatement des codages pratiqués. Un travail de *normalisation* est en cours pour y remédier . Cette normalisation sépare représentation physique et représentation logique des documents. Elle propose des conventions générales pour les différents types de textes.

38.1 Représentations logiques : SGML

Le Petit Robert fournit l'entrée suivante pour *linguistique* :

[phonétique] n.f. et adj. – 1826 ; de *linguiste*.

I N. f. 1 vx Etude comparative et historique des langues (grammaire comparée, philologie comparée). 2 (fin XIX^{e}) MOD. Science qui a pour objet l'étude du langage envisagé comme système de signes. " *La linguistique a pour unique [...] objet la langue envisagée en elle-même et pour elle-même* " (Saussure). [...]

II Adj. (1832) 1 Relatif à la linguistique. *Etudes linguistiques, Théories linguistiques*. => **distributionnalisme**, **génératif** (grammaire générative), **structuralisme**. 2 Propre à la langue, envisagé du point de vue de la langue. *Fait linguistique* => **langagier**. – *Expression linguistique*. *Signe*, *système, changement linguistique*. – *Communauté*, *géographie linguistique*. *Politique linguistique*. 3 Relatif à l'apprentissage des langues étrangères. *Vacances, séjours linguistiques à l'étranger*. – *Bain* linguistique*.

Cette entrée de dictionnaire fournit au lecteur humain de multiples indices lui permettant de classer les informations : le gras signale les renvois à d'autres entrées, les caractères droits les définitions et les renseignements techniques (datation, catégorie syntaxique ...). Les informations occupent une place relativement fixe : la transcription phonétique est au tout début, entre crochets, les datations après la catégorie, ou en début de définition. C'est une interprétation qui s'appuie sur la tradition lexicographique et les conventions propres à chaque dictionnaire. Les italiques servent à la fois à l'étymon (*linguiste*) et aux expressions utilisant le mot dans un de ses sens (avec des mises en facteur : *signe*, *système*, *changement linguistique*).

Les outils d'annotion, pour pouvoir utiliser un tel dictionnaire, doivent disposer d'un accès aisé aux différents types d'information. Le simple texte, même avec ses indications de présentation (gras, italiques, maigre, etc.), n'est pas directement utilisable. La représentation physique doit faire place à une représentation logique[224]. C'est l'équivalent de la transformation que nous avons opérée lors de la présentation de l'étiquetage lorsque nous avons

[224] N. Ide et J. Véronis (1995b) analysent en détail le codage des dictionnaires.

remplacé les notations positionnelles par une explicitation des types d'information (dans une structure trait-valeur).

Le balisage logique d'un document revient à indiquer sa structure : ses subdivisions et leurs relations. Il se réalise en deux étapes. La première est l'identification des éléments possibles pour un texte donné et de leurs relations. C'est en quelque sorte écrire une « grammaire de texte ». C'est ce qu'on appelle une Définition de Type de Document (DTD). La deuxième étape est l'introduction des balises choisies dans le document relevant de cette DTD, en respectant les règles éditées pour leur combinaison.

En adaptant au français la « grammaire de dictionnaires » fournie par N. Ide et J. Véronis (1995b) et en simplifiant à l'extrême, on peut distinguer les éléments suivants : la forme, subdivisé en orthographe et phonétique, et les homographes, relevant de parties du discours distinctes (*linguistique* {nom} et *linguistique* {adjectif}) et subdivisés en sens distincts :

entree ⌫ forme homographe+ | forme sens+[225]

forme ⌫ orthographe phonetique

homographe ⌫ categorie sens+

Chaque élément est encadré par deux balises de même nom, l'une ouvrante, l'autre fermante. Les balises sont entre chevrons. La balise fermante commence par une oblique. Le balisage concret serait alors :

```
<entree>
  <forme>
    <orthographe>linguistique</orthographe>
    <phonetique>à mettre</phonetique>
  <forme>
  <homographe>
    <categorie>nom</categorie>
    [...]
  <homographe>
    <categorie>adjectif</categorie>
        <sens>relatif à la linguistique</sens>
        <sens>propre à la langue, envisagé du point de vue de la langue</sens>
        <sens>relatif à l'apprentissage des langues</sens>
```

[225] Le signe + signifie que le constituant doit figurer au moins une fois et qu'il peut se présenter un nombre indéfini de fois.
La barre verticale sépare deux manières possibles de construire une entrée : une forme suivie d'homographes, ou une forme suivie d'un ou de plusieurs sens.
Une entrée de dictionnaire qui ne contiendrait pas d'indications orthographiques et phonétiques serait mal formée, par exemple.

```
    </homographe>
</entree>
```

Le balisage employé ici rend explicite ce qui n'existait que sous forme d'indices dans la version papier de l'entrée. Il obéit au langage standard de balisage SGML[226] qui est maintenant présent dans pratiquement tout logiciel de gestion de document[227]. SGML offre en plus des mécanismes particuliers pour noter les caractères « exotiques » en faisant abstraction de leur réalisation physique sur telle ou telle architecture. C'est le cas des caractères accentués, mais aussi de l'alphabet phonétique international. On peut ajouter de nouvelles conventions de notation pour les caractères ou suites de caractères non prévus, ce qui permet de faire face au caractère « ouvert » des notations nécessaires. Soulignons que SGML n'est pas une grammaire des textes possibles, mais un méta-langage permettant de définir la grammaire des différents types de textes[228].

38.2 Les types de textes : TEI

Une fois ce balisage logique introduit, il est possible d'accéder aux éléments d'information. On peut extraire la représentation phonétique (l'empan de texte compris entre <phonetique> et </phonetique>) ou les catégories des différents homographes ou les sens de l'adjectif, etc.

Ce premier niveau de normalisation s'avère cependant insuffisant. La grammaire complète définie peut suffire pour *Le Petit Robert*, elle peut se révéler inadaptée pour d'autres dictionnaires. En outre, rien n'empêche plusieurs groupes ou individus de se donner des conventions différentes pour un même type de document, ce qui empêche de comparer et d'échanger les résultats.

Un deuxième niveau est donc nécessaire. S'entendre sur des descriptions génériques pour les grands types de documents utilisés : dictionnaires, poésie, théâtre, oral, textes alignés, documents historiques, ainsi que pour les niveaux d'annotation qui peuvent les décorer : étiquettes, arbres, apparat critique, références croisées. Une initiative de grande ampleur, la TEI[229] *(Text*

---

[226] L'ISO (Organisation Internationale de Normalisation) a adopté en octobre 1986 SGML (Standard Generalized Markup Language) dans le but d'atteindre une réelle souplesse d'utilisation, de réutilisation et d'échange de l'information. Cette norme internationale (ISO 8879) a été rapidement adoptée par de nombreuses institutions privées et publiques, dans le monde anglo-saxon (American Association of Publishers, British Library, Oxford University Press, industrie aéronautique : Bœing, Airbus ...) mais aussi en France (Syndicat National de l'Edition, Cercle de la Librairie ...).

[227] Le succès grandissant de SGML tient aussi au fait qu'une grammaire particulière, HTML, issue de SGML décrit le langage hypertextuel utilisé pour le Web. Un traitement de texte courant, Word, offre ainsi la possibilité d'exporter un document en mode HTML.

[228] (van Herwijnen, 1994) constitue une introduction globale et pratique à SGML.

[229] Soutenue par l'Association for Computers and the Humanities, l'Association for Computational Linguistics et l'Association for Literary and Linguistic Computing. Le projet a été en partie financé par le National Endowment for the Humanities américain, la DG XIII de

Encoding Initiative) a depuis dix ans rassemblé des chercheurs de différentes disciplines et de toutes nationalités pour proposer des conventions sur ces types de documents. Elle a débouché sur des Recommandations[230] en 1994. De nombreux projets de constitution de corpus et de ressources linguistiques ont adopté la TEI (***BNC*** par exemple)[231]. Pour reprendre les termes de J. André (1996, p. 17), la TEI constitue un « inventaire – une sorte de flore, au sens de Buffon – des divers éléments pouvant constituer un document littéraire », et elle représente en ce sens une avancée dans la description et la formalisation des types de documents en circulation dans les diverses communautés langagières. Elle fournit ainsi indirectement des éléments pour les typologies de textes et les études sur les genres discursifs.

Il ne faut pas s'inquiéter de la lourdeur de ces balisages, dont témoigne l'exemple choisi. Ils ne sont absolument pas faits pour être insérés et utilisés « à la main ». Des environnements spécifiques permettent le balisage de textes et la vérification de la conformité du balisage effectué avec une « grammaire » fournie, tout comme les traitements de texte « cachent » à l'utilisateur les codages permettant de mémoriser la présentation qu'il a choisie.

39. DOCUMENTER UN CORPUS

Sans une documentation jointe, un corpus est mort-né. L'un des dangers de la facilité actuelle à rassembler des textes électroniques est précisément que les objectifs du regroupement ainsi que ceux des annotations effectuées ne soient pas enregistrés : le corpus cesse d'être utilisable dès que se perd la mémoire de ces choix.

La documentation doit couvrir deux volets distincts : les sources utilisées et la responsabilité éditoriale de constitution du corpus d'une part, les conventions d'annotation d'autre part[232].

---

la CEE, la fondation Andrew W. Mellon et le Social Science and Humanities Research Council du Canada.

[230] La TEI est donc une proposition de norme et non une norme.

[231] On trouvera dans (Ide et Véronis, 1995a) une présentation générale de SGML et de TEI, ainsi que les propositions relatives aux différents types de texte. Les *Cahiers Gutenberg* n° 24 (juin 1996) traduisent certains de ces articles et complètent l'information sur TEI et SGML.

[232] ***Susanne*** là encore est exemplaire : un livre entier (Sampson, 1995) informe sur ces deux volets du corpus, mais une documentation déjà très précise – reprise dans (Sampson, 1994) – est également fournie avec la version électronique. La TEI a fait des propositions détaillées sur le type de documentation à fournir pour un corpus (Dunlop, 1995).

39.1 Origine et histoire du corpus

L'information sur ce point doit indiquer les sources primaires utilisées, avec les références bibliographiques précises pour les éditions utilisées quand il s'agit de documents imprimés, mais aussi les objectifs visés par le regroupement, ses responsables, ainsi que les révisions qu'a subies le corpus au fil de sa mise au point.

39.2 Jurisprudence d'annotation

La qualité primordiale d'un système d'annotation, c'est sa cohérence interne[233]. Comme utilisateur d'un corpus annoté, on peut regretter tel ou tel choix. Par exemple, dans ***Susanne***, les deuxième, troisième, etc., éléments conjoints par une coordination sont représentés comme des subordonnés du premier (Sampson, 1994, p. 184). Une coordination de la forme *a, b and c* est indiquée ainsi *[a, [b], [and c]].* L'essentiel est que l'on puisse tabler sur la cohérence de traitement : toutes les coordinations sont effectivement notées ainsi. Si l'on s'intéresse à la coordination, on pourra filtrer les sous-arbres pertinents : leur forme globale ne varie pas. D'où l'importance des contrôles de qualité et des procédures de comparaison plus ou moins automatisés des résultats de plusieurs annotateurs / correcteurs sur les mêmes textes. Pour les 800 000 mots décorés syntaxiquement à l'université de Lancaster, le dispositif était le suivant. D'abord la double analyse pour comparer le travail d'un annotateur avec celui des autres : « Le but de la double analyse n'est pas tant la production d'un fragment correct que la détection de divergences significatives dans les pratiques d'annotation des deux analystes » (Black *et al.*, 1993, p. 34). Un logiciel permet de comparer les résultats de deux analystes sur un même texte. Il sert aussi aux analystes débutants à vérifier la qualité de leur travail au regard des annotations d'analystes plus chevronnés. Enfin, un grammairien expérimenté effectue une vérification approfondie par échantillonnage sur 1 % du résultat. Il importe également de contrôler la cohérence d'un annotateur au cours du temps[234] parce que sa compréhension des conventions d'annotation et sa finesse d'analyse évoluent.

Un corpus n'est compréhensible que si l'on dispose non seulement des étiquettes utilisées pour les mots comme pour les constituants, mais surtout d'informations sur le mode d'attribution de ces étiquettes et les critères de découpage sous-jacents : listes pour les catégories fermées, critères aussi précis que possibles pour les catégories ouvertes, assortis d'exemples, en

[233] C. Muller (1973, p. 10) le disait déjà voici longtemps, en particulier pour la segmentation et la lemmatisation.
[234] Nous ne connaissons pas d'études sur ce point. Cette absence s'explique sans doute par la difficulté à faire réanalyser les mêmes données à intervalles de temps suffisamment éloignés ou à trouver des données différentes présentant les mêmes difficultés d'annotation.

particulier des cas litigieux. Parallèlement aux corpus annotés, se développent, pour chaque schéma d'annotation, des guides d'annotation (*guidelines*), qui sont parfois plus justement dénommés des « recueils de jurisprudence » (*caselaws*). Si les découpages et la catégorisation n'ont en effet rien d'une science, il importe par contre de fixer la jurisprudence, à partir des décisions qui ont été prises dans tel ou tel cas, et qui éclairent ou rectifient les principes généraux qui ont été retenus. Les comparaisons de doubles analyses, en dehors des variations mineures, permettent de les établir. C'est la démarche suivie à Lancaster : « [...] les divergences importantes sont résolues par discussion (ou par appel à un tiers quand les deux analystes ne parviennent pas à un accord) » (Black *et al.*, 1984, p. 34). L'objectif de telles jurisprudences est d'assurer, dans la mesure du possible, une certaine reproductibilité de l'annotation : une compréhension solide de ces conventions doit permettre en principe à plusieurs analystes d'aboutir à une annotation la plus homogène possible.

L'expérience de Lancaster semble montrer, d'ailleurs, que l'annotation (ici sur le plan syntaxique, mais le propos peut être généralisé) ne peut pas reposer directement sur l'intuition, non étayée, des locuteurs, contrairement à ce qui avait été essayé dans une première phase. « [Les] annotateurs jouissaient d'une telle latitude dans les décisions à prendre lors de l'analyse manuelle qu'ils aboutissaient à un degré très bas de comparabilité des analyses. Plus intéressant, ils se sentaient mal à l'aise : avec si peu d'indications sur ce qui était ‘juste’ ou ‘faux’, ils se consultaient les uns les autres et développaient leur propre ‘norme’ non écrite sur la manière d'analyser les phrases, ou bien consultaient les traitements fournis dans les grammaires usuelles. Les conventions tacites et aléatoires développées ainsi pouvaient même être mutuellement incompatibles. Nous avons fini par céder à la demande de 'standards' de codification et le manuel d'analyse est devenu de plus en plus détaillé, jusqu'à réduire à un minimum les zones d'incertitude » (Black *et al.*, 1993, p. 41).

40. CONTRAINTES ET CONDITIONS INSTITUTIONNELLES

40.1 Assises institutionnelles

Comme nous l'avons vu pour les corpus étiquetés, il y a toujours à adapter une annotation donnée (changement de catégories, rajout de balises ...), soit pour comparer des annotations distinctes sur un même texte, soit pour ajouter, supprimer ou changer des catégories. Cela suppose d'abord des environnements informatiques adaptés : dans l'immédiat, ils sont créés au coup par coup et ne sont pas standardisés. Cela implique également une

identification fine des transformations et de leur difficulté, ce qui nécessite une certaine culture théorique et pratique issue de la tradition informatique des langages formels. Par exemple, nous l'avons vu, une notation dépendancielle ne se laisse pas forcément traduire en arbres.

Autant dire qu'une coopération approfondie entre informaticiens (spécialistes du TALN) et linguistes est nécessaire et le restera longtemps. Il semble d'ailleurs que le monde anglo-saxon arrive plus facilement à faire coopérer sciences humaines et sciences plus « dures », comme le montrent les conditions de réalisation de ***BNC*** ou de ***Penn Treebank***, alors qu'en France, la division entre « lettres » et « sciences » reste extrêmement forte (ne serait-ce que par l'existence d'universités distinctes pour chaque secteur).

Enfin, la constitution de corpus est une entreprise de longue haleine et coûteuse. Elle suppose des moyens financiers et institutionnels lourds. Le consortium à l'origine de ***BNC*** est significatif à cet égard[235]. On note l'alliance de compétences universitaires en linguistique et en informatique et d'entreprises privées, en particulier d'éditeurs, ainsi que le soutien de la puissance publique.

40.2 Problèmes juridiques

Peu de corpus sont dans le domaine public sans condition aucune[236] : l'accès aux documents primaires comme le fait de disposer du regroupement de documents et de leur annotation sont soumises à des restrictions diverses.

La présence de données personnelles peut faire obstacle à la mise à disposition de la communauté. C'est le cas de ***Menelas***. Même anonymisé (les noms propres de personne et de lieux sont remplacés par des chaînes de caractères conventionnelles), ce corpus fournit des informations personnelles (âge, symptômes, traitements) qui permettraient éventuellement de retrouver les patients concernés, violant ainsi le droit dont ils jouissent sur les informations les concernant (loi *Informatique et Libertés*).

L'attention s'est souvent centrée sur la protection des auteurs et ayant-droits des documents primaires (les ouvrages inclus dans un corpus). La protection de ceux qui ont annoté le corpus n'est pas moins importante. L'enrichissement d'un corpus par étiquetage ou parsage constitue en effet une plus-value considérable pour la recherche : il peut servir de base à de nouvelles annotations (apprentissage de chaînes de Markov ou de grammaires probabilistes). Les corpus résultant le plus souvent de la coopération de diverses personnes physiques et morales, il faut identifier

---

[235] Oxford University Press, Longman Group Ltd, Chambers Harrap, Oxford University Computing Services, Unit for Computer Research on the English Language (Lancaster University), British Library Research and Development Department. Ont par ailleurs contribué au financement de ce projet : UK Department of Trade and Industry, le Science and Engineering Research Council, ainsi que la British Library et la British Academy.

[236] À l'exception, notable, de ***Susanne***, déchargeable par ftp anonyme (Sampson, 1994, p. 187) : black.ox.ac.uk (ota/suzanne).

précisément les différentes parties prenantes et leurs droits.

Les interrogations juridiques peuvent donc concerner la création du corpus, sa protection une fois constitué et enfin sa diffusion[237]. Lors de la création du corpus, il s'agit d'abord d'identifier les « matériaux » visés et le régime juridique de chacun d'eux (certains peuvent être protégés par le droit d'auteur, d'autres non, comme fréquemment les textes officiels d'origine législative, administrative ou judiciaire, pour faciliter leur diffusion). Des autorisations, en fonction des traitements envisagés, peuvent être à demander non seulement pour le respect du droit pécuniaire et patrimonial mais aussi pour celui du droit moral[238] de l'auteur sur son œuvre (droit de divulgation, droit au respect de l'œuvre, etc.). La reproduction opérée peut en outre correspondre à un régime d'exception au droit de reproduction (usage privé, reproduction par des établissements de recherche, etc.). L'utilisation prévue du corpus influe aussi sur la nature des autorisations à négocier. Les produits issus d'un corpus (index, thesaurus, lexique) doivent également être protégés, au même titre que le corpus électronique lui-même. La diffusion du corpus peut se faire par cessions de droits, soit par licences d'utilisation (commercialisation par CD-ROM) soit par contrats d'abonnement ou d'interrogation.

[237] Le rapport de N. Pujol (1993) ne donne pas l'ensemble des situations qui peuvent se présenter et des attitudes à adopter, mais fournit une liste aussi exhaustive que possible des questions juridiques à se poser lors de la constitution d'un corpus, en particulier dans un cadre international. Nous nous inspirons de ce travail dans ce paragraphe.

[238] « L'œuvre étant manipulée en tout sens, il conviendra de s'assurer qu'il n'est pas porté atteinte au droit moral de l'auteur. Ce droit peut être menacé : a) par la mauvaise qualité du traitement linguistique b) mais aussi du seul fait que le traitement linguistique opéré ne participe pas du mode de reproduction de l'œuvre autorisé par l'auteur » (Pujol, 1993, p. 14).

CHAPITRE VIII

ANNOTER UN CORPUS

Nous ne prétendons pas fournir ici une présentation exhaustive. L'éclatement des réalisations, dispersées dans les publications, l'évolution rapide des outils, les avancées théoriques et pratiques conduisent à un « instantané » fragmentaire. Il est en outre difficile de prévoir les tendances à moyen terme. Notre objectif est de donner une idée des grands axes ... et des difficultés.

Dans la tradition pragmatique anglo-saxonne, les publications concernant les corpus mentionnent souvent les coûts des différentes opérations nécessaires. Ces renseignements permettent de prendre la mesure des moyens à mobiliser pour disposer des corpus réellement adaptés aux recherches linguistiques. À l'échelle de la francophonie, ils donnent une idée de l'ampleur des efforts à fournir. Ces " coûts " sont cependant donnés à titre indicatif. Ils donnent un ordre de grandeur, ils n'autorisent pas vraiment des projections, des comparaisons. À chaque niveau, les types d'annotation diffèrent trop pour qu'une mise en parallèle soit aisée. Pour s'en tenir à l'étiquetage, la taille du jeu d'étiquettes peut changer du tout au tout le coût de la correction.

41. NETTOYAGE ET HOMOGENEISATION

La phase initiale de « nettoyage » et d'homogénéisation des textes collectés sous forme électronique est une étape souvent sous-estimée, alors qu'elle est cruciale. Dans certains cas, les textes à intégrer dans un corpus ont été frappés pour la circonstance : ils contiennent des fautes de frappe ou d'orthographe. Dans d'autres cas, ils sont issus d'une reconnaissance

optique : il faut restituer les mots qui ont été répartis entre deux lignes, corriger les erreurs typographiques. Il peut s'agir également de textes déjà saisis pour d'autres fins (bandes de composition de livres ou de journaux), le codage qui y figure doit être pris en compte, pour être transformé ou supprimé.

Nous ne connaissons pas d'étude spécifique sur les coûts de cette phase. Le compte-rendu du projet AVIATOR (Blackwell, 1993) permet néanmoins d'évaluer les difficultés rencontrées. L'objectif est ici de développer des filtres permettant de « nettoyer » du texte tout-venant pour étudier l'évolution presque au quotidien de l'anglais, dans la perspective d'un corpus de suivi (cf. chapitre VII). Deux millions et demi de mots, provenant du journal *The Times*, sont traités chaque mois. Le titre même de ce compte-rendu donne une idée de l'ampleur du problème : « Des données sales au langage propre ». Comme S. Blackwell le souligne (*ibid.*), la correction de ce qui semble être des erreurs typographiques ne va pas forcément de soi. Une orthographe non standard a parfois pour but d'imiter une prononciation étrangère, dialectale ou idiolectale. Ou bien le mot a été forgé dans une optique ludique[239] (mot-valise, déformations diverses). Il s'agit alors de choix délibérés de la part de l'énonciateur, qui doivent donc être conservés comme tels. Les données comprennent parfois des codes propres au traitement pour lequel les documents étaient destinés au départ (par exemple des indications de photocomposition). Les titres, sous-titres et légendes suscitent aussi un traitement spécifique : quoiqu'ils constituent des unités à part entière, à ne pas mêler au texte qui les environne, ils sont généralement dépourvus de ponctuation finale. Il faut donc distinguer leur début et leur fin.

42. SEGMENTATION

La segmentation consiste à découper une suite de caractères en « unités » : mots simples ou unités polylexicales.

42.1 Repérer les unités

Le repérage des « mots » est délicat[240]. Un certain nombre de caractères, en effet, fonctionnent tantôt comme séparateurs de mots tantôt comme composants de mots. C'est le cas du trait d'union, qui joint deux mots dans *vient-il*, mais pas dans *va-et-vient*[241]. C'est le cas encore de l'apostrophe :

[239] Cf. (Fiala et Habert, 1989 ; Renouf, 1993).
[240] (Silberztein, 1993, p. 111-136) montre la complexité des phénomènes.
[241] (Mathieu-Colas, 1994) montre l'hétérogénéité extrême des emplois du trait d'union dans les

séparateur comme guillemet simple, pour signaler l'élision, composant dans *aujourd'hui*, les abréviations et la représentation du langage parlé : *v'la au't chose*. C'est le cas surtout de l'espace, partie intégrante des unités complexes : *une carte bleue*.

Les unités complexes occupent une place importante en français. On estime au cinquième d'un texte la surface qu'elles couvrent. Pour le français, des inventaires extrêmement fournis ont été réalisés au LADL, sous l'impulsion de M. Gross, aboutissant à un dictionnaire électronique de « mots composés » ou DELAC (Courtois, 1990 ; Silberztein, 1993, p. 60-108). Ce dictionnaire associe aux séquences retenues des indications sur leurs variations éventuelles (flexion, discontinuités, alternances lexicales) ainsi que leurs propriétés syntaxiques (transformations[242]).

Mentionnons la difficulté à découper automatiquement le texte en phrases : titres, énumérations séparées par des points-virgules, exemples insérés dans le texte et faisant interposition, etc. La ponctuation offre des indices peu fiables[243]. Le point est une marque d'abréviation, un séparateur dans des codes (01.41.13.24.63) ou des nombres (3.13) , un indice d'alignement (dans une table des matières) et une fin de phrase. Or le découpage en phrases est crucial pour de nombreux traitements : examen des cooccurrences, étiquetage et analyse syntaxique ...

42.2 Techniques

Pour isoler les « mots », on écrit des règles qui emploient le contexte pour statuer sur les limites des unités. Par exemple, un trait d'union ayant à sa droite un pronom clitique comme *je*, *tu*, *il* a un statut de délimiteur. Il sépare un verbe de son pronom sujet conjoint (un *t* d'appui peut s'interposer). Ces règles sont combinées avec le recours à des dictionnaires de mots simples ou complexes (par exemple, comprenant la liste des mots français qui incluent en leur sein l'apostrophe, comme *aujourd'hui* ou *prud'hommes*).

Le système INTEX[244] (Silberztein, 1993) est l'exemple d'un segmenteur associant règles et dictionnaires. À partir des dictionnaires électroniques du LADL, il assure le découpage initial d'un texte tout-venant, l'étiquetage des mots simples et la reconnaissance des unités polylexicales. Son approche est basée sur des règles et non sur des probabilités. Il combine deux traitements : la projection sur le texte des dictionnaires, ce qui associe à chaque " mot " la ou les étiquette(s) pertinente(s) ainsi qu'aux suites de mots (éventuellement discontinues) leurs lectures éventuelles comme " mots composés " ou " expressions composées ", puis une désambiguïsation par

---

dictionnaires.

[242] Par exemple, *analyse des données* au sens statistique n'accepte pas le pluriel pour *analyse* ni le singulier pour *données* ni le remplacement de *des* par *de*.

[243] Pour le rôle de la ponctuation dans l'analyse syntaxique, voir (Nunberg, 1990).

[244] Les techniques éprouvées des automates et des transducteurs à états finis lui donnent une grande efficacité.

des " grammaires locales " (*ibid.* p. 154-167). Par exemple, la *phrase Luc a travaillé pour le Ministère de l'intérieur* admet deux interprétations (*ibid.*, p. 139) *: C'est de l'intérieur que Luc a travaillé pour le Ministère* et *C'est pour le Ministère de l'intérieur que Luc a travaillé*. Il y a conflit entre deux unités polylexicales : *Ministère de l'intérieur* et *de l'intérieur*. La représentation produite signale les deux découpages : Luc a travaillé pour le 1[Ministère 2[de l'intérieur]2]1 où les indices identifient les deux possibilités. En l'occurrence, l'ambiguïté n'est pas levée. Dans d'autres contextes, on peut trancher. Des « grammaires locales » élaguent le graphe que constitue le texte dans lequel ont été ajoutées les étiquettes des mots simples et les expressions et mots composés. Elles permettent d'éliminer certains chemins[245]. Par exemple, lorsqu'un mot peut être pronom clitique ou déterminant et qu'il est suivi d'une forme qui ne peut être qu'un verbe, comme dans : *Max le veut*, l'étiquette {pronom clitique} est éliminée.

42.3 Difficultés

Les unités polylexicales occupent une place fondamentale dans le lexique. Un segmenteur qui ne dispose pas d'inventaires de ces unités va « émietter » à tort les textes. De multiples techniques ont été testées pour faciliter le repérage automatique de ces mots complexes. Certaines d'entre elles ont été évoquées au chapitre II. D'autres reposent sur le filtrage statistique des mots qui « s'attirent » au sein d'un contexte restreint, d'autres encore sur l'utilisation de patrons syntaxiques (du type [{nom} {préposition} {nom}] comme cadre de vie), d'autres enfin combinent ces deux approches (Daille, 1993). Cependant, nombre de séquences proposées par ces outils ne constituent pas en fait des dénominations (cf. II 3.3)[246]. Les inventaires d'unités complexes réalisés pour le TALN suscitent généralement la perplexité ou la contestation sur la délimitation faite et sur le choix de considérer telle séquence comme une unité dénominative plutôt que comme un syntagme libre. Le risque symétrique de l'« émiettement » est de considérer à tort des suites de mots comme des unités polylexicales.

L'utilisation de dictionnaires comprenant un nombre important d'unités complexes fait naître en outre des ambiguïtés pour les séquences qui fonctionnent comme un tout dans certains domaines et qui sont à considérer comme des syntagmes libres dans d'autres. Dans « l'analyse des données montre que ... », le segment *analyse des données* peut renvoyer à une

[245] Soulignons l'extrême généralité du traitement effectué. Cela permet d'utiliser INTEX pour d'autres traitements : étiquetage sémantique etc.

[246] Cet excédent s'explique partiellement par le caractère encore fruste des techniques employées. Il tient plus fondamentalement aux limites de nos connaissances sur les mécanismes langagiers de création d'unités dénominatives. Les contraintes sémantiques à l'œuvre sont encore très peu explorées. Enfin, les dénominations possibles constituent un sur-ensemble des dénominations effectives, il n'est pas sûr qu'on puisse modéliser la manière dont une communauté langagière choisit au sein des dénominations possibles.

famille précise de techniques statistiques (présentée dans le chapitre IX), et c'est alors une unité, ou bien il doit être pris « au pied de la lettre », comme un groupe de mots sans lien particulier[247]. Plus les inventaires d'unités complexes s'étendent, plus ils rendent probables ces rencontres de hasard. Il n'est pas toujours sûr qu'il faille faire l'hypothèse, lorsqu'on rencontre une séquence inventoriée, de la présence effective de cette séquence.

43. ÉTIQUETAGE MORPHO-SYNTAXIQUE

Attribuer à chaque mot la ou les étiquettes possibles peut se faire par consultation d'un dictionnaire, où chaque forme est suivie d'une liste de catégories, ou par analyse morphologique, ou par combinaison des deux techniques. Pour lever l'ambiguïté, deux solutions, qui peuvent d'ailleurs être associées, s'offrent alors : le recours à des règles ou l'appel aux probabilités (ce qui est sans doute la tendance dominante)[248].

43.1 Taux d'ambiguïté

Il est nécessaire, pour évaluer la tâche d « désambiguïsation » morpho-syntaxique, c'est-à-dire le choix de l'étiquette correcte parmi les étiquettes possibles, d'évaluer le nombre moyen d'étiquettes pour un mot. M. El Bèze et T. Spriet (1995) donnent les informations suivantes : « [...] une très grosse part de l'ambiguïté syntaxique est détenue par un petit nombre de mots fréquents [..]. De plus, ces mots sont essentiellement des mots outils. Ils appartiennent à des classes fermées et jouent un rôle syntaxique bien cerné dans la littérature. » Ils précisent (*ibid.* p. 58) : " [...] 30 % de l'ambiguïté est détenue par les 8 mots ambigus les plus fréquents[249] (50 % par les 36 premiers) mais il faut traiter 1 825 formes différentes pour lever 90 % de l'ambiguïté[250]. » E. Tzoukermann *et al.* (1996) précisent ce premier constat sur deux ensembles de 94 882 et 200 182 occurrences respectivement, tous deux extraits du journal *Le Monde* (septembre-octobre 1989 et janvier 1990) :

Nombre d'étiquettes	% du corpus de 94 882 mots	% du corpus de 200 182 mots

[247] On ne sait pas attacher de manière fiable à une unité polylexicale une indication de domaine (*analyse de données* : mathématiques, statistiques) et encore moins s'en servir pour n'utiliser que les unités propres au domaine, d'autant que les domaines sont « perméables » : la linguistique peut recourir à l'expression *analyse des données* dans ses deux acceptions.
[248] J.-P. Chanod et P. Tapanainen (1995b) les comparent précisément, à partir d'une même segmentation et d'un même analyseur morphologique. Ils donnent l'avantage à l'approche par règles.
[249] Ces 8 formes sont : *la le l' les en un une a*.
[250] Les chiffres de J.-P. Chanod et P. Tapanainen (1995b) concordent globalement.

1	57 %	58 %
2	26 %	25 %
3	11 %	11 %
4	0,5 %	1 %
5	0,9 %	2 %
6	2 %	2 %
7	0,5 %	0,5 %
8	0,5 %	0,1 %

Plus de la moitié des mots ne soient pas ambigus. Le nombre de mots pouvant relever de 4 à 8 étiquettes est très restreint (4.4 % dans le premier cas, et 5.6 % dans le second). Le taux moyen d'ambiguïté par mot se monte alors à 1.72 pour le premier corpus et à 1.81 pour le second[251].

43.2 Désambiguïsation par règles

Certaines suites de catégories sont illicites. Par exemple, deux étiquettes sont possibles pour *le* {déterminant} ou {pronom} et pour *guide* {verbe} ou {nom}. Cependant, toute la combinatoire n'est pas réalisable dans la séquence *le guide*. Des quatre possibilités, seules sont actualisables [{Pronom} {verbe}] (*il le guide*) et [{déterminant} {nom}] *(le guide commence son exposé*). On peut donc écrire une première règle d'élagage qui remplace la combinatoire par les deux seules suites licites de catégories. On utilise alors des règles « négatives ».

D'autre part, certaines formes permettent d'édicter des règles « positives ». Elles imposent en effet des contraintes fortes sur celles qui les précèdent ou les suivent immédiatement. Ainsi, *me* ou *te* sont suivis soit d'un pronom clitique (*il me le donne*) puis d'un verbe soit directement d'un verbe. On peut alors s'appuyer sur cette information pour éliminer des ambiguïtés. Dans *il me le garde*, *le* ne peut être qu'un pronom clitique et *garde* qu'un verbe. De telles formes servent de levier pour désambiguïser une partie de leur entourage. On parle d'« îlots de confiance ». Les clitiques post-posés et reliés par un trait d'union offrent également de tels appuis (dans *Route-t-il correctement le courrier*, *route* ne peut être qu'un verbe). Les formes nouvellement désambiguïsées servent à leur tour de point d'appui : les îlots de confiance vont croissant.

Les outils de désambiguïsation sont donc de manière générale des « grammaires locales » (Silberztein, 1993) qui prennent en entrée le graphe correspondant à la projection des différentes étiquettes sur le texte et éliminent une partie des chemins de ce graphe, ou inversement qui rajoutent des chemins (par exemple pour rendre compte des unités complexes comme

[251] M. El-Bèze et T. Spriet (1995, p. 52-53) donnent des chiffres proches.

bien que ou *carte bleue*)[252]. Les automates ou transducteurs correspondants ne savent pas traiter les dépendances à longue distance que l'on trouve en syntaxe. C'est également le cas en désambiguïsation probabiliste.

43.3 Désambiguïsation probabiliste

La désambiguïsation probabiliste s'appuie sur le caractère positionnel de langues comme le français et l'anglais, lequel fournit des contraintes locales fortes. Dans le graphe orienté des étiquettes possibles pour chacun des mots, il s'agit de chercher le chemin de probabilité maximale. Le choix de l'étiquette la plus probable en un point donné se fait au regard de l'historique des dernières étiquettes qui viennent d'être attribuées. En général, cet historique se limite aux deux ou trois étiquettes précédentes, on parle alors de bigrammes ou de trigrammes. Il repose sur des chaînes de Markov (Calliope, 1989, p. 360-370, Mérialdo, 1995, p. 11-13).

Ces méthodes supposent de disposer d'un corpus d'apprentissage. Ce corpus d'apprentissage doit être d'une taille suffisante pour permettre une estimation fiable des probabilités des suites de catégories et des différentes catégories d'un mot donné dans ces enchaînements. Le coût de préparation de ce corpus d'apprentissage est important. On procède alors par approximation. Un premier corpus d'apprentissage, relativement court, permet d'étiqueter un corpus plus important. Celui-ci est corrigé, ce qui permet de réestimer les probabilités. Il sert donc à un second apprentissage. Et ainsi de suite.

Les unités polylexicales sont mal prises en compte dans cette approche. Ainsi, pour reprendre l'exemple de M. El-Bèze et T. Spriet (1995), les adjectifs et participes placés immédiatement à droite du nom composé *cour d'appel* s'accordent avec *cour* et non avec *appel*. La probabilité d'un adjectif ou d'un participe passé féminin singulier après un nom masculin singulier comme *appel* sera pourtant donnée comme très faible par le corpus d'apprentissage, à juste titre d'ailleurs. Plus généralement, les désambiguïsations qui reposent sur un contexte large échappent à ce type de méthode. Des ambiguïtés comme première / troisième personne du singulier dans *je ne le pense pas* / *il ne le pense pas* ne sont pas éliminées, parce que ces étiqueteurs probabilistes s'appuient sur le contexte de la catégorie précédente, voire des deux catégories précédentes, pour trancher, et qu'ici il faudrait prendre en compte les trois catégories précédentes (Chanod et Tapanainen, 1995a).

L'approche probabiliste suppose par ailleurs que le corpus d'apprentissage ne présente pas des fonctionnements langagiers trop différents du corpus à étiqueter. Dans le cas de ***BNC***, un certain nombre de mots comme *I*, *well* et

[252] J.-P. Chanod et P. Tapanainen (1995b) ont ainsi développé un étiqueteur qui comprend 75 règles. E. Tzoukermann *et al.* (1995) donnent des exemples des règles qu'ils ont mises au point pour le français.

right etaient mal étiquetés dans la partie orale du corpus dans la mesure où l'apprentissage avait été réalisé sur la partie écrite (Leech *et al.,* 1994).

43.4 Performances

Aucun dictionnaire ne peut être entièrement exhaustif. En outre, les entrées du dictionnaire peuvent être incomplètes (certaines catégories, pourtant possibles, en sont omises). Un analyseur morphologique ne fournit pas non plus d'hypothèses sur la totalité des mots à étiqueter. Il reste donc toujours des « mots inconnus », ne serait-ce qu'en raison des noms propres, des mots empruntés à des langues étrangères ou des néologismes (*débureaucratiser*).

Les taux habituellement cités tournent autour de 95 à 98 % d'étiquettes justes. Ce chiffre paraît encourageant. Cependant, ces performances incluent souvent les ponctuations parmi les formes étiquetées. Or les ponctuations couvrent environ 10 à 15 % de la surface des textes, ce qui diminue d'autant le nombre des formes lexicales qui sont effectivement correctement catégorisées. Par ailleurs, nous l'avons vu, une bonne moitié des formes d'un texte ne relève que d'une catégorie et d'une seule. La désambiguïsation est donc à comptabiliser sur le reliquat seulement, ce qui double le pourcentage d'erreur. Notons enfin que 5 % d'erreur, c'est une étiquette erronée tous les 20 mots, soit plus d'une fois par phrase dans un texte courant. Une telle « performance » handicape un parseur intervenant en aval.

La fiabilité d'un étiqueteur donné est à évaluer à l'aune des tâches qui vont avoir recours par la suite au texte étiqueté : les enjeux ne sont pas les mêmes s'il s'agit d'analyse syntaxique automatique ou d'étude de la répartition de certains patrons morpho-syntaxiques. Il convient aussi de comparer les résultats affichés avec ceux qui proviennent d'une intervention manuelle. M. Marcus *et al.* (1993) indiquent : « l'étiquetage manuel a pris à peu près deux fois plus de temps que la correction d'un étiquetage automatique, avec un taux de désaccord entre personnes étiquetant à peu près double, et un taux d'erreur presque de 50 % plus élevé. »

Il est en outre extrêmement difficile de comparer les performances : les jeux d'étiquettes, leur taille changent d'un système à l'autre : 37 catégories pour (Chanod et Tapanainen, 1995), 253 pour Tzoukermann *et al.,* 1995) par exemple. Le taux d'ambiguïté d'un étiquetage est en effet proportionnel à la taille du jeu d'étiquettes employé. Il faut également tenir compte de la stabilité des résultats : si le taux d’ambiguïté restant ne varie que faiblement (1.2 %) dans les expériences d'E. Tzoukermann *et al.* (1995) selon qu'ils emploient un jeu de 67 ou de 253 catégories, 2.5 % des formes ont été analysées différemment, (Stein et Schmid, 1995, p. 29), des résultats relativement divergents sont donc fournis. En outre, les ambiguïtés possibles ne sont pas de même nature : on ne peut mettre sur le même plan l'hésitation entre nom et verbe (*porte*) et celle entre adjectif et participe passé. Dans ce cas, la levée d'ambiguïté n'a pas les mêmes conséquences pour les traitements

ultérieurs : considérer un mot comme adjectif ou participe passé changera peu la place qui lui sera attribuée dans la structure construite.

43.5 Post-traitement et coûts

Pour un usage linguistique fin, le post-traitement manuel s'avère en tout cas indispensable. Malgré les environnements spécialisés qui ont été développés, la correction reste coûteuse. Dans le cadre de ***BNC***, elle est évaluée (Leech *et al.,* 1994), après le passage d'un étiqueteur probabiliste (CLAWS4, basé sur les chaînes de Markov), au taux de succès de 96 à 97 %, à 40 minutes de travail spécialisé pour 1 000 mots, soit 41 années-homme pour 100 millions de mots. Il faut en outre prendre en compte le nombre d'étiquettes : plus il est grand, plus il rend difficile la correction manuelle. Cette difficulté pousserait à choisir des étiquettes « connues », basées sur le savoir grammatical courant (sur la terminologie grammaticale traditionnelle), pour faciliter le travail des correcteurs et l'utilisation ultérieure par des chercheurs (Greenbaum, 1993).

Pour le corpus de l'université de Lancaster, près de 39 minutes (Black *et al.,* 1994, p. 60) sont nécessaires au traitement de 1 000 mots (pré-traitement, passage de l'étiqueteur probabiliste CLAWS, correction manuelle).

43.6 Evaluation et nouvelles tendances

Eric Brill (1995) résume ainsi les points forts et les faiblesses des deux approches : « [Les] étiqueteurs stochastiques ont bien des avantages sur les étiqueteurs bâtis manuellement, en particulier ils rendent superflue la construction laborieuse de règles manuelles, et saisissent des informations utiles qui peuvent ne pas avoir été remarquées par l'analyste humain. Cependant, les étiqueteurs stochastiques présentent l'inconvénient que les connaissances linguistiques ne sont capturées qu'indirectement, par le biais de grands tableaux statistiques. »

L'écriture de règles se heurte rapidement à la complexité des interactions effectives entre les règles. En effet, chaque règle agit sur un texte qui a été modifié par les règles précédentes. Il faut donc prévoir autant que faire se peut ces interactions, qui peuvent devenir d'une complexité très grande, voire ne plus être maîtrisables. À l'inverse, la mise au point des règles peut s'appuyer sur l'intuition des locuteurs.

L'étiquetage et la désambiguïsation, comme d'autres secteurs de l'annotation des données textuelles, donnent lieu à des approches mixtes, où un étiquetage probabiliste est corrigé *in fine* par des règles du type de celles

évoquées ci-dessus, ou vice-versa[253].

Les techniques d'apprentissage sont également mises à contribution. La tentative la plus achevée est actuellement celle d'E. Brill (1995), dont l'étiqueteur est en cours d'adaptation pour le français. Le système dispose d'un dictionnaire associant aux formes les probabilités qu'elles portent telle ou telle catégorie. La catégorie la plus probable est projetée sur le corpus de mise au point. Les erreurs commises ainsi sont repérées par comparaison avec la version étiquetée à la main de ce corpus. Le système propose des règles de correction, assez proches finalement de celles qui ont été évoquées ci-dessus. Elles sont de la forme : changer une étiquette *a* en étiquette *b* si le mot précédent est étiqueté *w*. Elles prennent en compte un contexte étroit : deux positions avant ou après la forme examinée. Sont retenues les règles qui améliorent le plus l'état de la catégorisation, c'est-à-dire qui enlèvent le plus d'erreurs et en ajoutent le moins. Ces règles sont alors appliquées. Une nouvelle comparaison et une nouvelle génération et application de règles sont opérées, jusqu'à ce qu'il ne soit plus possible de corriger le texte sans ajouter davantage d'erreurs qu'on n'en corrige. C'est une autre forme, automatique cette fois, du processus mentionné de « tâche d'huile » autour d'îlots de confiance. E. Brill indique par exemple que son système « apprend » 447 transformations sur un corpus d'entraînement de 600 000 mots avec une exactitude de 97.2 %, mais que les 100 premières suffisent à assurer une désambiguïsation exacte à 96.8 % (*ibid.*, p. 557). Ces règles peuvent s'appuyer soit sur les catégories, éventuellement multiples, soit aussi sur les mots dominés par les catégories.

Pour reprendre les termes de Leech et de ses collègues (1994, p. 61) : « La guerre contre l'erreur est [...] une guerre d'usure, dans laquelle des stratégies variées sont employées, mais où il ne faut pas s'attendre à une solution-miracle. Le rôle de la personne qui corrige *a posteriori* reste crucial, mais l'élimination de l'erreur est une tâche qui est, petit à petit, passée à l'ordinateur. »

44. ANALYSE SYNTAXIQUE

Nous mentionnons avant tout l'analyse syntaxique automatique. L'analyse syntaxique manuelle nécessite surtout de disposer d'un environnement informatique facilitant la tâche de parenthésage et de catégorisation des constituants. Elle rend plus cruciale la vérification de l'homogénéité des

[253] Comme l'indiquent M. El-Bèze et T. Spriet (1995, p. 48) : " [...] il suffit d'écrire 4 à 5 règles pour traiter environ 50 % des erreurs commises par un système probabiliste. " E. Tzoukermann *et al.* (1995) constituent comme autant de modules un analyseur morphologique, un ensemble de règles d'élagage et un étiqueteur probabiliste : ils les combinent de diverses manières (en retenant 43 possibilités, jouant sur des seuils et des ordres distincts) et examinent les performances selon les choix, ce qui les conduit à utiliser d'abord les règles puis les probabilités.

résultats.

44.1 Structuration par règles

44.1.1 Règles « négatives »

On retrouve pour le parsage une technique déjà utilisée pour l'étiquetage : l'élagage (*pruning*). Il s'agit dans le domaine syntaxique d'utiliser des règles « négatives », qui ont pour fonction d'éliminer les hypothèses non justifiées. C'est l'approche du parseur ENCG, ce qui amène Voutilainen et Heikkila (1994, p. 190) à parler d'analyseur « réductionniste ». Pour chaque étiquette morphologique d'un mot donné, les fonctionnements syntaxiques possibles sont fournis. Par exemple, un nom peut être sujet, objet, complément prépositionnel, etc. L'élagage élimine les fonctionnements illégitimes en contexte. Ces contraintes syntaxiques (400 dans le cas présent) sont elles-mêmes issues d'études intensives de corpus (Karlsson, 1994, p. 122). En principe, ces règles d'élagage sont indépendantes les unes des autres et n'ont pas besoin d'être ordonnées. Il semblerait cependant qu'une grammaire ENCG reste assez « fragile ».

44.1.2 Règles " positives "

Elles peuvent être de complexité plus ou moins grande. Les grammaires à affixes du projet TOSCA (Nederhof et Koster, 1993, p. 166-170) qui décorent des règles hors contexte d'affixes représentant des paramètres, des attributs ou des traits, permettent une grande finesse de comportement : vérification des accords et des compatibilités sémantiques etc.

44.2 Structuration probabiliste

Les parseurs reposant sur des règles butent sur deux types de problèmes, comme le rappelle M. Rajman (1995, p. 158) : la couverture linguistique et l'ambiguïté. Couverture : les règles mises au point sont soit trop permissives (elles acceptent des énoncés incorrects) soit au contraire trop restrictives (elles refusent des agencements de mots pourtant valides). Ambiguïté : le nombre d'hypothèses proposées est souvent très important (cf. chapitre II).

L'idée générale du parsage probabiliste[254] est de remplacer la distinction

[254] (Rajman, 1995) fournit une introduction générale aux modèles probabilistes pour l'analyse syntaxique. (Black *et al.*, 1993) constitue une présentation beaucoup plus détaillée, à la fois en ce qui concerne l'apprentissage des paramètres d'un modèle probabiliste et pour l'interaction entre approche

binaire acceptable / non acceptable pour un couple <séquence, structure> par une probabilité, les séquences inacceptables pouvant correspondre alors à une probabilité nulle (*ibid.* p. 159). Les deux problèmes mentionnés trouvent là leur solution. Certains agencements sont reconnus comme rares, mais possibles. D'autres prennent une place centrale, leur probabilité étant forte. La probabilité attribuée à chaque structure pour une phrase donnée permet de classer les structures par probabilité croissante, et de garder la ou les structures de plus forte probabilité. Un corpus arboré de départ sert à l'apprentissage du modèle : la probabilité des différentes réalisations d'un syntagme donné est estimée à partir de sa fréquence dans ce corpus[255]. L'utilisation du modèle sur un corpus plus large permet de vérifier l'adéquation du modèle et de l'améliorer (en accroissant le corpus d'apprentissage).

44.3 Performances et évaluation

Puisque, nous l'avons vu, l'annotation syntaxique peut varier énormément en complexité, il est malaisé de comparer les résultats de différents parseurs. Une des possibilités, encore peu explorée (Atwell *et al.,* 1994), consiste à « aligner » plusieurs représentations syntaxiques d'un même texte. Une version rudimentaire de cette approche (Black *et al.,* 1993, p. 4) consiste à réduire l'annotation aux parenthésages, en éliminant toutes les étiquettes, pour ne garder donc que les découpages structurels et leurs emboîtements. On peut alors aisément comparer deux parenthésages et repérer les désaccords. C'est ce qui est appelé (*ibid.*) le « score de cohérence structurelle » (*structural consistency score*). Une autre optique consiste à soumettre un ensemble de phrases de test à plusieurs analyseurs et à comparer, avant tout manuellement, leurs résultats. Cette deuxième démarche sert plutôt à examiner de manière fine les réactions des parseurs : chaque phrase est centrée autour d'un phénomène syntaxique bien défini, elle est donc souvent relativement simple par rapport aux énoncés effectivement rencontrés par les parseurs dédiés au texte tout venant. On manque en tout état de cause de données comparatives.

Un premier critère d'évaluation est celui de la justesse linguistique des résultats retenus. Elle est difficile à apprécier. On peut tout de même opposer des analyseurs (et partant des corpus arborés) qui visent à un simple dégrossissage et ceux qui, au prix éventuellement d'un post-traitement important, aboutissent à des analyses vérifiées et cohérentes au sein du cadre théorique choisi et qui peuvent servir de pierre de touche à des

---

par règles et analyse probabiliste. Ce livre résulte d'une collaboration étroite, pendant cinq ans, entre le centre de recherche IBM Watson et l'université de Lancaster (UCREL - Unit for Computer Research on the English Language).

[255] En principe, ce corpus doit être aussi vaste que le permettent les moyens rassemblés. La précision des estimations qu'il autorise en dépend. La collaboration IBM Watson - Université de Lancaster a abouti par exemple à l'analyse manuelle de 800 000 mots (Black *et al.,* 1993, p. 16).

recherches linguistiques fines. Pour le système TOSCA, H. van Halteren et N. Oostdijk (1993, p. 155) indiquent que, pour les textes de fiction, dans 88 % des cas, l'analyse juste fait partie des résultats produits par le parseur, alors que cette proportion tombe à 56 % pour les textes qui ne relèvent pas de la fiction. Malheureusement, ils ne fournissent pas d'hypothèses sur les raisons de ce décalage. Les textes « informatifs » comprennent-ils des phrases plus longues, des constructions spécifiques (par exemple propres à des disciplines scientifiques ou techniques) qui ne se rencontreraient pas dans les textes de fiction ? Selon A. Voutilainen et J. Heikkila (1994, p. 194), le parseur ENCG donne l'étiquette syntaxique correcte d'un mot dans 96 % des cas (85 % environ des mots n'ont plus qu'une seule étiquette syntaxique à la fin du processus d'émondage, mais avec un taux d'erreur de 3 %). Les constats de (Black *et al.,* 1993, p.2-5), voici quelques années, sont plus sévères. Les auteurs parlent de « déplorable état de l'art » (*ibid.* p. 2) et citent trois expériences peu encourageantes. Dans la première, trois des auteurs chercheurs à IBM Watson ont procédé de manière indépendante, en 1990, à l'évaluation de quatre parseurs importants pour l'anglais, sur 35 phrases de 13 mots extraites au hasard de dépêches (2 millions de mots) de l'agence *Associated Press*. Les avis concordaient : un des systèmes analysait 60 % des phrases correctement. Les scores des trois autres parseurs allaient de 35 à 40 % de résultats justes. Deuxième expérience : en 1992, le concepteur d'un parseur important a pris 50 phrases de 13 mots dans ***Brown***, en variant les genres choisis. Il a indiqué les frontières de constituants à la main, préparant ainsi la « bonne réponse ». Il a ensuite utilisé son parseur : les résultats étaient corrects dans 30 % des cas seulement. Troisième expérience : la comparaison en 1992 des résultats de sept parseurs sur 100 phrases de longueur variable (de 4 à 69 mots avec une moyenne de 22 mots) tirées au hasard d'un million de mots du *Wall Street Journal*. La correction moyenne du simple parenthésage (sans prendre en compte les étiquettes) ne dépassait pas 22 %, et les résultats s'étalaient de 16 % à 41 % de résultats structuralement corrects.

Un second critère d'appréciation, concernant les parseurs et les grammaires qu'ils utilisent, est la réutilisation possible ou effective de l'approche soit sur d'autres secteurs de la même langue soit pour d'autres langues. C'est ainsi que le parseur ENCG développé pour l'anglais a été adapté au suédois, au danois et au basque (Voutilainen et Heikkila, 1994, p. 191).

Un troisième critère, lié au précédent, mais plus difficile à apprécier, parce que moins factuel, est celui de la " coloration théorique " des conventions d'annotation. À quel cadre théorique sous-jacent renvoient-elles ? Notons tout de même que la tendance est plutôt, sinon à des notations consensuelles, ce qui n'a pas grand sens, du moins à des pratiques évitant les distinctions controversées et les parti-pris méthodologiques trop marqués[256]. C'est

[256] Une exception au moins : le corpus de 65 000 mots d'oral transcrit (enfants de 6 à 12 ans) analysé manuellement (Polytechnic of Wales) qui s'inspire étroitement de la Grammaire Fonctionnelle Systémique de Halliday.

nécessaire pour que le corpus puisse être réutilisé (Black *et al.,* 1993, p. 37).

Il est enfin un critère que nous écarterons, celui du temps nécessaire au parsage lui-même[257]. D'abord parce qu'il est difficile de donner des informations comparables (les langages informatiques utilisés, la taille des mémoires, leur configuration changent notablement le sens des mesures). Ensuite parce le temps de calcul n'est plus une ressource rare, et qu'en outre l'amélioration des performances des machines le réduit continuement. Enfin, parce que l'optimisation des parseurs est un art fructueux[258], mais qu'il faut probablement attendre une plus grande maturité du domaine pour qu'elle soit vraiment à l'ordre du jour pour les corpus arborés.

44.4 Post-traitement et coûts

C'est la phase de « nettoyage » manuel des résultats fournis par le parseur utilisé.

Il peut s'agir, comme pour le système TOSCA, de choisir entre les analyses alternatives proposées (Halteren et Oostdijk, 1993, p. 157-159). Sont utilisées des forêts partagées (*shared forests*), qui mettent en facteur commun les sous-arbres partagés. L'annotateur examine la phrase en contexte et sélectionne à chaque point d'ambiguïté le sous-arbre approprié.

A l'inverse, dans le cas de ***Penn Treebank***, où le parseur déterministe *Fidditch* (Hindle, 1994), fournit une analyse syntaxique unique pour chaque phrase, mais laisse des constituants non rattachés, la tâche des annotateurs est d'attacher les constituants « orphelins ». Voici pour la phrase *Battle-tested industrial managers here always buck up nervous newcomers with the tale of the first of their countrymen to visit Mexico, a boatload of warriors blown ashore*, l'état des traitements fourni dans (Marcus *et al.,* 1993, p. 322-325) :

1) Analyse syntaxique automatique produite par *Fidditch* :

Les constituants non attachés débutent par ?. Les syntagmes prépositionnels commençant par *of* sont attachés à un nom s'ils en suivent un (c'est le cas pour *tale of*, *boatload of*), et restent non attachés dans le cas contraire (*first of*). Les virgules, qui peuvent jouer le rôle de conjonctions, fragmentent aussi l'ensemble d'arbres.

```
( (S
  (NP (NBAR (ADJP (ADJ "Battle-tested/JJ")
                  (ADJ "industrial/JJ"))
            (NPL "managers/NNS")))
```

[257] A titre anecdotique, deux chiffres, empruntés à Hindle (1994, p. 116) : avec Fidditch, de l'ordre de 6 heures pour analyser un million de mots, et presque deux semaines pour analyser 44 millions de mots de dépêches de l'agence *Associated Press*.

[258] F. Karlsson indique ainsi (1994, p. 142) qu'une réécriture du parseur ENCG a fait passer le temps d'analyse de 3 à 5 mots seconde à 400 à 500 mots seconde ...

```
(? (ADV "here/RB"))
(? (ADV "always/RB"))
(AUX (TNS *))
(VP (VPRES "buck/VBP"))
(? (PP (PRES "up/RP")
     (NP (NBAR (ADJ "nervous/JJ")
          (NPL "newcomers/NNS")))))
(? (PP (PREP "with/IN")
     (NP (DART "the/DT")
       (NBAR (N "tale/NN")
         (PP of/PREP
           (NP (DART "the/DT")
             (NBAR (ADJP
               (ADJ "first/JJ")))))))))
(? (PP of/PREP
     (NP (PROS "their/PP\$")
       (NBAR (NPL "countrymen/NNS")))))
(? (S (NP (PRO *)
       (AUX to/TNS)
       (VP (V "visit/VB")
         (NP (PNP "Mexico/NNP"))))))
(? (MID ",/,"))
(? (NP (IART "a/DT")
     (NBAR (N "boatload/NN")
       (PP of/PREP
         (NP (NBAR
           (NPL "warriors/NNS"))))
       (VP (VPPRT "blown/VBN")
         (? (ADV "ashore/RB"))
         (NP (NBAR (CARD "375/CD")
           (NPL "years/NNS")))))))
(? (ADV "ago/RB"))
(? (FIN "./.")))
```

2) Après simplification automatique et avant correction manuelle :

La représentation est simplifiée pour faciliter la tâche des annotateurs en rendant le résultat visuellement plus clair et en éliminant des distinctions mineures (nom propre / nom commun, par exemple).

```
( (S
  (NP (ADJ Battle-tested industrial)
      managers)
  (? here)
  (? always)
  (VP buck)
  (? (PP up
       (NP nervous newcomers)))
  (? (PP with
       (NP the tale
         (PP of
           (NP the
             (ADJP first))))))
  (? (PP of
       (NP their countrymen)))
  (? (S (NP *)
      to
      (VP visit
        (NP Mexico))))
  (? ,)
  (? (NP a boatload
       (PP of
         (NP warriors))
    (VP blown
       (? ashore)
       (NP 375 years))))
  (? ago)
  (? .))
```

3) Après correction manuelle :

L'environnement utilisé permet d'attacher un constituant, de changer sa position dans l'arbre, de modifier son étiquette ... Grâce à des notations

spécifiques, on peut d'une part indiquer qu'une séquence est un constituant majeur mais que sa catégorie syntaxique est sujette à discussion, et d'autre part rendre compte des ambiguïtés réelles : c'est le cas pour *blown ashore 375 years ago* qui peut modifier soit *warriors* soit *boatload*, d'où l'indication *pseudo-attach*.

```
( (S
  (NP Battle-tested industrial managers
     here)
  always
  (VP buck
     up
     (NP nervous newcomers)
     (PP with
        (NP the tale
           (PP of
              (NP (NP the
                    (ADJP first
                       (PP of
                          (NP their countrymen)))
                    (S (NP *)
                      to
                      (VP visit
                         (NP Mexico))))
                  ,
                  (NP (NP a boatload
                        (PP of
                           (NP (NP warriors)
                              (VP-1 blown
                                 ashore
                                 (ADVP (NP 375 years)
                                    ago)))))
                    (VP-1 *pseudo-attach*))))))))
.)
```

44.5 Coûts

Pour l'insertion manuelle d'arbres syntaxiques rudimentaires (parenthésage et étiquetage des constituants), la vitesse peut atteindre une phrase par minute (Black *et al.,* 1993, p. 20). La moyenne pour l'analyse syntaxique manuelle effectuée à l'université de Lancaster est de 51 minutes pour 1 000 mots : cela comprend pré-traitement, parenthésage et étiquetage grossier dans un environnement informatique spécifique et post-traitement (*ibid.* p. 60).

D'après (Marcus *et al.,* 1993, p. 323), la correction des résultats du parseur utilisé pour ***Penn Treebank*** suppose un temps d'apprentissage (de l'ordre de deux mois) plus long que le nettoyage de l'étiquetage. La vitesse moyenne de correction est alors de l'ordre de 475 mots l'heure (voire 575 ou 675 quand les sorties du parseur sont simplifiées avant correction). L'évaluation faite est la suivante (*ibid.*) : « À un taux moyen de 750 mots par heure, une équipe d'annotateurs à temps partiel travaillant 3 heures par jour devrait arriver à 2,5 millions de phrases analysées corrigées en un an, chaque phrase étant corrigée une seule fois. »

Il faut en outre prévoir le temps de familiarisation avec les conventions d'annotation syntaxique. (Black *et al.,* 1993) indique ainsi qu'il a fallu attendre six mois d'apprentissage en moyenne avant que le travail d'un annotateur devienne optimal.

44.6 Difficultés

Tout ne ressortit pas à un format d'arbre. C'est le cas des éléments parenthétiques qui forment des structures autonomes, non reliées au reste de la phrase. Cela suppose que le parseur puisse suspendre l'analyse englobante, effectuer celle d'un tel élément, et reprendre l'analyse de plus haut niveau (Briscoe, 1994, p. 98). À supposer que l'on arrive à analyser automatiquement de telles structures, il reste à disposer des notations adéquates.

La distinction entre les arguments d'un verbe et ses simples modifieurs s'avère extrêmement délicate à ajouter de manière cohérente. Le dessein, ***dans Penn Treebank***, était d'ajouter manuellement cette information. La difficulté rencontrée a conduit à faire machine arrière. De la même manière, ***Susanne*** n'a pas réussi, malgré des efforts soutenus des annotateurs, à intégrer un classement des compléments en termes de grammaire de cas, à la Fillmore : « la nature des relations logiques que des prédicats variés entretiennent dans l'usage réel avec leurs arguments s'est avérée trop diverse pour un tel traitement, et l'équipe croit avoir 'testé jusqu'à épuisement'[259] l'hypothèse selon laquelle la structure propositionnelle de

[259] *tested to destruction*

base en anglais peut être adéquatement décrite grâce à un ensemble limité de 'cas » (Sampson, 1994, p. 185). Les relations entre les pronoms et leurs antécédents n'ont pas non plus été ajoutées à ***Susanne***, probablement moins par peur de déboucher sur des apories que faute de moyens.

Toute grammaire « fuit », pour reprendre une image souvent employée dans la communauté du parsage robuste. L'idée de rendre compte de l'ensemble des phénomènes syntaxiques de la langue (on parle de la « couverture » de la grammaire utilisée par un parseur) est un fantasme, stimulant certes, comme tous les mythes, mais illusoire, comme le soulignent du point de vue linguistique J.-M. Marandin (1993) et du point de vue du TALN T. Briscoe (1994, p. 100). Une raison de fond : la langue varie. Dans le temps d'abord. Mais aussi selon les genres discursifs et les domaines d'emploi. À la différence des langages formels utilisés en logique ou en informatique, l'ensemble des règles n'est pas donc fini. Ce constat, classique pour le lexique, soulève plus de réticences en syntaxe.

45. ÉTIQUETAGE SEMANTIQUE

L'une des grandes méthodes d'analyse sémantique de corpus suppose des connaissances préalables et consiste à projeter ces connaissances sur le corpus pour en faire ressortir certaines propriétés. C'est sur ce principe que repose le travail de M. Sussna (1993) et la plupart des recherches en matière de désambiguïsation lexicale.

Le principe général de cette méthode est simple. On étiquette le corpus pour l'enrichir d'informations sémantiques. Pour ce faire, on exploite généralement des données lexicales et non contextuelles, connaissances générales sur les sens d'un mot, le concept ou le thème auquel il renvoie. Ceci permet alors d'observer le fonctionnement du mot en contexte. De multiples expériences ont été menées dans cette optique[260] : elles diffèrent par le jeu d'étiquettes utilisé et par la méthode d'étiquetage.

Toutefois, les données lexicales initiales font parfois défaut. C'est même souvent le cas lorsque le corpus à traiter relève d'une langue spécialisée. Il faut alors commencer par construire les catégories sémantiques devant servir à étiqueter le corpus.

[260] Une variante de cette méthode consiste à projeter des connaissances non pas sous la forme d'étiquettes destinées à enrichir le texte, mais sous la forme de patrons qui permettent de sélectionner de manière ciblée des données considérées comme pertinentes. Nous ne développons pas cet aspect ici. (Hearst, 1992) exploite, par exemple, cette méthode pour rechercher des relations hyponymiques dans un corpus destinées à enrichir un thesaurus existant.

45.1 Construire des catégories sémantiques

La difficulté de réutiliser les bases lexicales spécialisées, l'inadéquation des bases lexicales générales et plus fondamentalement le manque de ressources lexicales, notamment pour le français (cf. chapitre III), soulèvent la question de l'acquisition des connaissances lexicales. La construction manuelle de ce type de base de données requiert l'expérience d'un lexicographe et, pour les langues spécialisées, celle d'un expert du domaine. Le coût et la difficulté de ces entreprises ont mis à l'honneur les méthodes automatiques ou semi-automatiques qui considèrent les corpus comme des sources de connaissances pour la construction de catégories sémantiques, dans l'idée qu'elles puissent servir ensuite à étiqueter des corpus.

La construction de ces catégories sémantiques — qu'il s'agisse de classes de synonymes, de groupes de mots relevant d'un même champ sémantique ou d'un même thème — suit toujours le même principe général. La démarche consiste à :

- définir le contexte d'un mot, de manière à identifier les mots qui cooccurrent avec lui, l'ensemble des mots qui figurent dans le même contexte et qui, dans une approche distributionnelle de la sémantique en décrivent le sens ;
- définir une mesure de similarité entre les mots deux à deux, chaque mot étant représenté par les relations de cooccurrence dans lesquelles ils entrent ;
- exploiter cette mesure de similarité pour construire des classes de mots considérés comme équivalents selon le point de vue considéré (par exemple, des synonymes ou des mots relevant du même domaine…).

À ces trois étapes correspondent trois « ordres d'affinité » (Grefenstette, 1994b), trois niveaux de relations entre les mots[261] : les relations de cooccurrence, de similarité et d'équivalence[262]. Le travail de G. Grefenstette présenté au chapitre IV suit cette démarche générale. Nous nous appuyons sur cet exemple dans ce qui suit.

45.1.1 Définir un contexte

Le choix de la nature du contexte dépend du corpus exploité et des relations sémantiques recherchées. G. Grefenstette retient le syntagme nominal pour identifier les noms sémantiquement voisins et le document pour construire les familles de mots (cf ; chapitre IV, section 2). Trois grandes classes de

[261] Nous ne considérons ici que les relations entre mots, mais les affinités peuvent être calculées pour d'autres unités : on a vu (en III-2) que G. Grefenstette calcule des similarités entre des expressions, en l'occurrence des groupes nominaux (1993).

[262] Nous généralisons le propos de G. Grefenstette en décrivant le troisième ordre d'affinité comme celui des relations d'équivalence plutôt que comme celui des axes sémantiques qui nous semblent avoir un statut intermédiaire entre la similarité et l'équivalence.

contextes peuvent être identifiées : les contextes graphiques, syntaxiques et documentaires. L'extrait de ***Menelas*** suivant montre la différence, pour le mot *épisode*, entre une fenêtre de 7 mots (encadrée) et le contexte syntaxique tel que le définit (Grefenstette, 1994) (en italiques) :

> Depuis cette époque on ne note aucune récidive d'angor jusqu'à il y a
> 8 jours où il a *présenté* un *épisode* de *précordialgie survenant* à l'effort, durant
> environ 45 minutes, sans irradiation[263].

Les contextes graphiques se définissent comme des fenêtres de mots : deux mots cooccurrent s'ils figurent à moins de *x* mots de distance[264] dans l'ordre linéaire du texte. La taille de la fenêtre dépend des relations sémantiques que l'on recherche, les cooccurrences à petite, moyenne et grande distance tendant respectivement à faire ressortir des expressions figées ou semi-figées (*prendre pour*, *avoir faim*), des contraintes de sélection (*boire* / *vin*) et des mots appartenant au même champ sémantique (Lafon, 1981; Church et Hanks, 1990). Le calcul des fenêtres graphiques ne nécessitant qu'un corpus segmenté, elles sont souvent privilégiées pour le traitement de gros corpus.

L'apparition de corpus arborés permet désormais de définir des contextes syntaxiques. Seuls les mots appartenant au même syntagme ou, mieux, en relation de dépendance syntaxique sont alors retenus comme cooccurrents. Pour étudier les contraintes de sélection, on considère ainsi les relations sujet-verbe ou verbe-objet (Church et Hanks, 1990 ; Hindle, 1990) tandis qu'on prend le groupe nominal comme contexte pour repérer les classes d'adjectifs (Assadi et Bourrigault, 1995). Cette approche syntaxique suppose de disposer d'un corpus arboré ou partiellement arboré et généralement désambiguïsé sur le plan morpho-syntaxique[265], mais elle engendre moins de bruit que l'approche graphique[266] : les contextes linguistiquement aberrants (l'association *jours – épisode* dans l'exemple ci-dessus) sont éliminés. Cela rend cette approche bien adaptée aux corpus de taille moyenne (Basili *et al.*, 1993a ; Bouaud *et al.,* 1997).

Les contextes documentaires, enfin, sont définis à partir d'une unité textuelle (paragraphe, partie, article, chapitre, document…). C'est ce type de contexte que G. Grefenstette définit pour le calcul des variantes.

De nombreux auteurs ne retiennent par ailleurs que les contextes les plus significatifs. Ce filtrage *a posteriori* des contextes préalablement extraits est le

---

[263] Nous n'avons pas considéré ici que les groupes prépositionnels *durant 45 minutes* et *sans irradiation* devaient être rattachés à *épisode*. Pour l'anglais, G Grefenstette résout le problème du rattachement du groupe prépositionnel par des règles *ad hoc* (1994).

[264] En général, les relations de cooccurrence ne sont pas orientées et l'ordre dans lequel figurent les mots est indifférent.

[265] On peut toutefois proposer des méthodes de pondération des analyses concurrentes en cas d'ambiguïté syntaxique. Voir par exemple (Grishman et Sterling, 1994).

[266] « [N]on seulement les associations syntaxiques reflètent une information fonctionnelle, ce que ne font pas les paires rapprochées sur une base graphique, mais la méthode d'extraction de ces associations syntaxiques est aussi *plus efficace*, le nombre d'associations utiles détectées étant considérablement plus élevé que ce qu'on obtient par des méthodes reposant sur une distance graphique. » (Basili *et al.*, 1993a, p. 154). L'analyse syntaxique fonctionne en effet comme un premier filtre.

plus souvent statistique[267] : on ne retient comme cooccurrents que les mots figurant « anormalement » souvent dans les mêmes contextes[268].

45.1.2 Calculer des similarités

Une fois définie la notion de contexte, on peut calculer pour un mot l'ensemble de ses cooccurrents, sa distribution. Cette distribution sert alors à représenter les mots et permet de les comparer entre eux. C'est l'approche suivie par G. Grefenstette et décrite au chapitre IV. Concrètement, cela signifie qu'un mot se représente par un vecteur sur l'ensemble des cooccurrents possibles, *i.e.* sur l'ensemble des mots du corpus. La similarité entre deux mots est mesurée comme une distance entre les vecteurs représentant chacun de ces mots[269].

Ces mesures de similarités sont difficiles à exploiter en tant que telles. Les scores obtenus ne s'interprètent pas dans l'absolu mais seulement relativement les uns aux autres. Par ailleurs, les mesures ou les classements obtenus résistent à l'interprétation. On a souvent besoin de savoir sur quels critères deux mots sont rapprochés

Le problème vient plus fondamentalement de ce qu'une liste triée des similaires d'un mot donné n'est pas une classe : ces listes sont centrées autour d'un mot pôle et ce n'est pas parce que *ship* (*navire*) et *truck* (*camion)*, par exemple, sont tous les deux similaires à *boat* (*bateau*) (Hindle, 1990) que les deux relations de similarités sont comparables ni que *ship* et *truck* sont nécessairement similaires entre eux. Partant de ce constat, G. Grefenstette (1994) propose de structurer cette liste des similaires d'un mot selon ses différents axes sémantiques, ce qui revient à distinguer différents types de similarités. J. Bouaud et ses collègues (1997) choisissent de représenter un ensemble de relations de similarités sous la forme d'un graphe qui situe un mot dans un réseau de similarités et fait ressortir des zones denses, riches en similarités croisées. Pour aller plus loin dans cette voie, il faut construire des classes sémantiques à partir d'une relation d'équivalence entre les mots. C'est là pour nous le véritable troisième ordre d'affinité.

---

[267] Ce n'est cependant pas le seul type de filtrage possible : pour la recherche de collocations, F. Smadja (1993) filtre les collocations sur une base syntaxique, ou même en fonction de leur degré de figement.

[268] Voir par exemple (Lafon, 1981), (Church et Hanks, 1990) ou (Justeson et Katz, 1996). D'autres auteurs, visant la construction de classes sémantiques plutôt que la recherche de collocations, considèrent au contraire que le seul fait qu'un contexte soit attesté une fois suffit à le rendre significatif (Bensch et Savitch, 1995 ; Bouaud, 1997). Signalons par ailleurs qu'un filtrage statistique ne peut s'effectuer que sur un volume important de données.

[269] Nous préférons parler ici de similarité entre les mots plutôt que de distance comme le font les travaux de classification automatique. Le terme de distance sémantique est d'ordinaire employé pour désigner des distances calculées à partir d'une taxonomie ou d'un réseau (cf. *supra*). G. Grefenstette (1994) ou P. Bensch et W. Savitch (1995) s'inspirent de la mesure de Jaccard ou Tanimoto mais la littérature sur les méthodes de classification présente de multiples mesures de similarité (Saporta, 1990 ; Lebart et Salem, 1994) et différentes mesures sont employées en acquisition de connaissances sémantiques.

45.1.3 Construire des classes de mots

Cette étape n'est pas abordée dans le traitement lexicographique de G. Grefenstette (1993), mais cette piste est explorée par d'autres auteurs, pour la modélisation d'un domaine, notamment[270]. En interprétant le score de similarité entre les mots comme une mesure de distance entre des objets, on peut appliquer les méthodes de classification automatique pour construire des classes de mots. Il s'avère cependant que les classes induites à partir de corpus sont difficiles à exploiter. Les méthodes purement inductives produisent des regroupements de mots hétérogènes. Pour construire des catégories sémantiques cohérentes, il faut corriger ces premiers résultats en fusionnant ou en scindant certaines classes pour obtenir une granularité régulière, en éliminant les intrus, parfois en reconstituant « à la main » des classes complètement éclatées.

Pourtant, si l'on considère l'ampleur et la difficulté de la tâche consistant à donner une description lexicale de l'ensemble des mots d'un corpus, et d'un corpus spécialisé notamment, il s'avère que les connaissances lexicales induites à partir de corpus, aussi bruitées et imparfaites soient-elles, sont précieuses. Ce sont des ébauches qui proposent une première organisation du matériau lexical et permettent d'amorcer le travail de description. A. Mikheev et S. Finch (1995) soulignent par exemple l'intérêt de ces méthodes de classification pour la modélisation des connaissances d'un domaine : « [l]a construction de classes sémantiques de mots à partir de corpus permet au cogniticien de repérer les principales catégories ou principaux types sémantiques existant dans le domaine en question et d'organiser le lexique en regard de ces types. ».

45.1.4 Procéder par itérations

La construction de catégories sémantiques repose généralement sur une alternance d'induction de connaissances à partir de corpus et d'interprétation, *i.e.* de projection de connaissances extérieures au corpus. Une première classification permet d'identifier une ou plusieurs classes cohérentes qui peuvent être figées puis projetées sur le corpus sous la forme d'un étiquetage partiel. Seuls les mots de ces premières classes porteront une étiquette de classe, mais ils constituent des îlots de confiance à partir desquels une nouvelle classification peut être construite[271]. Cette méthode incrémentale est donc une méthode mixte consistant à induire des connaissances même

[270] Voir, entre autres, (Assadi et Bourrigault, 1995), (Bensch et Savitch, 1995), (Mikheev et Finch, 1995), (MacMahon et Smith, 1994) ou (Bouaud *et al*., 1997).

[271] C'est la démarche adoptée par Bouaud *et al.* (1997) ou P. Bensch et W. Savitch (1995, p. 12) : « quand on applique notre technique de classification […] à un corpus réel, elle identifie un ensemble de catégories qui paraissent naturelles, sans toutefois classer beaucoup de mots dans ces catégories. Mais, il s'est avéré que ce petit nombre de mots classifiés dans un premier temps pouvait servir de point de départ pour classifier d'autres mots. ».

parcellaires que l'on peut ensuite projeter sur le corpus pour en induire de nouvelles.

Une variante de cette démarche incrémentale part non des premières classes induites mais d'un étiquetage grossier du corpus. C'est ce que font R. Basili *et al*. (1993b) ou R. Grishman et J. Sterling (1994) mais aussi Z. Harris (voir chapitre VII).

45.2 Projeter des catégories sur un corpus

45.2.1 Segmentation en unités sémantiques

Déjà présente au niveau morpho-syntaxique, la question de la segmentation du corpus se pose d'autant plus au niveau sémantique que la tradition fait davantage défaut. Quelle unité de sens faut-il retenir ? On considère souvent le mot, par solution de facilité parce que les sources lexicales utilisées sont elles-mêmes structurées autour des mots, aux expressions polylexicales et mots composés près. Dans certains cas, cependant, les unités inférieures sont à étiqueter : pour une étude thématique de ***Enfants***, les préfixes négatifs doivent être comptés au même titre que les adverbes de négation, lesquels comportent au contraire généralement plusieurs mots (*ne… pas*). Il est par ailleurs souvent difficile d'identifier les mots qui, dans un syntagme ou dans une phrase, doivent porter une étiquette donnée. Dans ***Enfants***, les expressions *difficultés financières*, *pas assez d'argent*, *considérations financières* ont toute une connotation négative, mais à quel mot associer cette étiquette négative ?

45.2.2 Désambiguïsation sémantique

Si les problèmes d'ambiguïté sont négligés — dans la langue de spécialité notamment —, l'étiquetage peut se faire hors contexte, sur la liste des formes du texte. C'est l'approche de (Basili *et al*., 1993c) semble-t-il. Pourtant, l'objectif est généralement de désambiguïser le corpus et l'étiquetage doit être fait en contexte.

L'étiquetage manuel est envisageable pour les corpus de taille moyenne (en deçà du million de mots) s'il faut choisir parmi quelques étiquettes générales parce que les cas ambigus sont rares et faciles à trancher : « Une fois qu'une classe sémantique est clairement définie, avec l'aide d'une interface conviviale, l'étiquetage à la main d'un mot est l'affaire de quelques secondes. Nous avons résolu de simplement sauter les mots pour lesquels le choix d'une étiquette n'est pas évident[272] ou pour lequel aucune étiquette ne

[272] C'est-à-dire s'il prend plus de « 30 secondes » (Basili *et al*., 1993a) (NDA).

paraît adaptée.» (*ibid.*, p. 346-347). « On n'a pas forcément besoin de faire appel à un linguiste pour l'étiquetage, [même si] on a besoin d'un linguiste pour établir un jeu d'étiquettes approprié. » (Basili *et al*., 1993a, p. 157).

S'il faut procéder à un étiquetage fin en revanche, la procédure manuelle devient sujette à erreur, difficile à homogénéiser et surtout trop coûteuse. « [L]a partie du corpus Brown qui est étiquetée par les classes de mots de WordNet, un exemple de corpus important , disponible et désambiguïsé à la main, montre clairement combien il est difficile d'obtenir des données 'satisfaisantes'. Ce corpus est relativement petit (de l'ordre de quelques centaines de milliers de mots) en comparaison de la taille des corpus actuels (plusieurs millions ou dizaines de millions de mots) ; la méthode d'annotation qui a été utilisée est très coûteuse en temps de travail […] ; et la qualité des résultats reflète la difficulté de la tâches standards actuels (les annotateurs sont en désaccord dans environ 10% des cas […]). » (Resnik, 1995).

D'où le besoin de méthodes automatiques robustes de désambiguïsation de corpus et l'intérêt des travaux qui, comme (Sussna, 1993), cherchent à les mettre au point.

CHAPITRE IX

QUANTIFIER LES FAITS LANGAGIERS

Divers outils informatiques permettent d'extraire, à partir de corpus ayant fait l'objet d'un travail d'annotation, les occurrences d'unités textuelles qui correspondent à un patron donné (mot, lemme, catégorie grammaticale ou sémantique, patron syntaxique, etc.). Ces outils permettent aisément de constituer la liste exhaustive des contextes où cette unité-pôle apparaît. L'examen des différents contextes d'une unité textuelle projette un éclairage indispensable sur les emplois que cette unité trouve dans le corpus, faisant apparaître des régularités qu'une lecture cursive du corpus n'aurait pas toujours révélées. Cependant dès que le nombre des contextes est un peu élevé, les mises en contextes ainsi réalisées (comme les concordances, etc.) deviennent des objets difficilement manipulables, même sous forme informatisée. L'organisation de ces listes (définition et ordre de présentation des contextes) influence très fortement la perception de divers phénomènes relatifs à la forme-pôle.

Le tableau 1 regroupe quelques lignes extraites des 5 030 contextes de la forme *je* dans ***Mitterrand1***. Ces contextes sont triés par ordre alphabétique, d'après la forme qui suit le pôle. Une telle approche permet de remarquer, en inspectant l'ensemble des lignes de contexte réalisées pour cette forme, que les occurrences de *je* sont prises dans des répétitions plus longues: *je le crois, je le dis*, etc.

Tableau 1. — Extrait d'une concordance de la forme je dans ***Mitterrand1***

ue la france qui a acquis,	je le crois,	la confiance et le respect
ères personnels, aussi, et	je le crois,	qui se réfèrent à la moral
cer des propositions pour,	je le crois,	saisir le monde entier du
rté des facilités qui ont,	je le crois,	sauvé le secteur du textil

```
ation de la fin du siècle. je le crois tout à fait, sans quoi je n
n souvient aussi- cela est je le crois, tout à fait, venu de consi
de la république: je suis, je le crois, très fidèle à ce que je su
jours, j' ai observé avec, je le crois, une grande patience, pour
ants que cela contribuera, je le crois, utilement au redressement
bre de plans, j' ai donné- je le crois vraiment- plus d' expansion
rachever le portrait. moi, je le dessine tous les jours, par des a
ite, je l' ai dit à alger, je le dirai à amman en jordanie où je s
 dans le monde. la france, je le dirai simplement, a déjà apporté
```

Pour généraliser ce type de démarche à l'ensemble des formes du corpus, il faut mettre en oeuvre des procédures de quantification qui éviteront au chercheur d'avoir à examiner l'ensemble des contextes de chacune des formes du corpus.

Ce chapitre propose un survol des approches quantitatives les plus courantes d'un corpus de textes[273]. La section 1 présente des objectifs de recherche qui conduisent à opérer des décomptes textuels à des fins de comparaison. Les problèmes liés à l'identification des unités dans le texte sont abordés dans la section 2. La section 3 traite du repérage des séquences d'unités. Les sections 4 et 5 introduisent ensuite des méthodes permettant de comparer les décomptes réalisés au sein d'un corpus partitionné. La section 6 est consacrée à l'articulation des décomptes réalisés à partir de différents systèmes d'annotation. Nous terminons (section 7) par un exemple de recherche sur les séries textuelles chronologiques qui combine plusieurs des méthodes présentées dans le chapitre.

46. POURQUOI QUANTIFIER ?

Au-delà des études centrées chaque fois sur un type d'unité textuelle particulier, s'est développé un courant dont les dénominations ont varié au cours du temps[274], et qui se fixe pour but l'étude quantitative des faits langagiers. L'approche quantitative permet seule d'accéder à la description de phénomènes textuels qui présentent un grand intérêt une fois mis en évidence et dont il aurait été difficile de cerner les contours *a priori*..

46.1 Étudier la variation de traits linguistiques dans un corpus

Certaines études menées par des linguistes se fixent pour but principal la description de la variation, au sein d'un corpus, de l'ensemble des éléments d'un même système d'unités linguistiques (graphèmes, formes, lemmes, lexies, système de catégories grammaticales, séquences, etc.). En général, ce type de tâche s'accommode mal de procédures de segmentation et

[273] Chacune de ces méthodes est présentée dans (Lebart et Salem, 1994).
[274] Cf., par exemple, (Herdan, 1964), (Muller, 1968).

d'identification approximatives des unités de décompte. Il nécessite au contraire que le texte analysé soit soumis, lors d'une étape préalable, à une réflexion minutieuse sur les procédures de repérage, d'identification et d'annotation des unités à recenser. Une fois les comptages réalisés pour chacune des unités du système, on soumet ces décomptes à des traitements statistiques afin de mettre en évidence les variations des différentes unités.

46.2 Réaliser des typologies de textes et de documents

Un courant relativement ancien de l'analyse quantitative des textes opère des quantifications dans le but de réaliser des typologies portant sur l'ensemble des textes réunis en corpus. Le problème de l'attribution d'auteur[275] en est un exemple. Il s'agit de déterminer si tel ou tel texte, sur lequel on manque de renseignements, présente des caractéristiques quantitatives laissant supposer qu'il a pu être écrit par un auteur dont on possède par ailleurs des échantillons de textes. On s'efforce donc de déterminer des systèmes d'unités discriminantes qui permettent de trancher en matière d'attribution. La comparaison des descriptions quantitatives des différents textes doit permettre dans ce cas d'obtenir des indications qui ne résultent pas de connaissances *a priori* sur les textes mais bien des similitudes qu'ils présentent au plan quantitatif.

On a recours à des méthodes comparables lorsqu'il s'agit de prélever parmi un vaste ensemble de documents ceux d'entre eux qui peuvent présenter de l'intérêt pour une tâche particulière (problème de l'indexation et de la récupération de documents industriels).

Pour ce second type d'études, le problème de la nature linguistique des unités qui permettent de mener à bien les tâches entreprises n'est pas central puisque le but ultime est le regroupement de textes. La sélection du système des unités de décompte qui sert de base aux comparaison se fait avant tout en fonction de l'efficacité pratique de l'ensemble de la démarche au regard de la tâche considérée.

Ces deux types de préoccupation (sections 1.2 et 1.2) se combinent parfois en proportions variables dans des études particulières. La mise en place de procédures à visées typologiques pose du même coup le problème du choix des unités les mieux à même de faire ressortir des oppositions.

46.3 Déceler des corrélations entre phénomènes

Une étude portant sur la répartition des pronoms personnels de la première

[275] Le travail de (Holmes, 1985) présente une revue assez complète des travaux en matière d'attribution d'auteur.

personne dans chacune des huit années de ***Mitterrand1*** montre que la fréquence d'emploi de ces pronoms varie sensiblement au cours du temps. On constate sur la figure 1, une tendance à l'augmentation du pronom *je* et une diminution du pronom *nous*. Cette tendance s'inverse légèrement dans la dernière année du septennat. Comme on le voit, les deux phénomènes manifestent une certaine liaison au cours du temps.

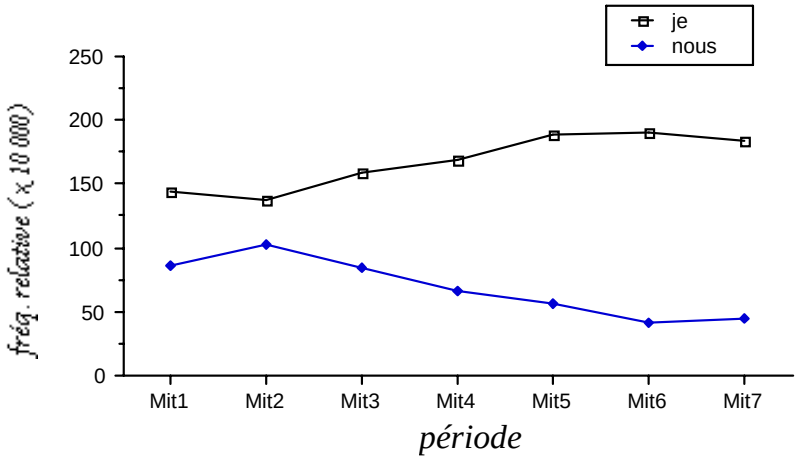


Figure 1.— Évolution des formes je et nous dans Mitterrand1[276]

On comprend aisément que ces variations de fréquences intéressent des spécialistes du texte politique. L'entrée quantitative est ici la seule voie d'accès à l'analyse détaillée et contrastive d'un tel phénomène.

47. LES UNITES

La méthode statistique s'appuie sur des mesures et des comptages réalisés à partir des objets que l'on veut étudier. Décompter des unités, les additionner entre elles, cela signifie, d'un certain point de vue, les considérer, au moins le temps d'une expérience, comme des occurrences identiques d'un même type. Pour soumettre une série d'objets à des comparaisons statistiques il faut donc, dans un premier temps, définir une série de liens systématiques entre des cas particuliers et des catégories plus générales.

Dans la pratique, l'application de ces principes généraux implique que soit définie une *norme de dépouillement* permettant d'isoler à partir du texte annoté les différentes unités sur lesquelles porteront les dénombrements.

[276] Le nombre des occurrences de chaque forme, dans chaque partie, est rapporté à la longueur de la partie considérée et multiplié par 10 000 pour une plus grande lisibilité des résultats.

Ch. Muller (1973) expose les difficultés liées à l'établissement d'une telle norme de dépouillement

> La norme devrait être acceptable à la fois pour le linguiste, pour ses auxiliaires, et pour le statisticien. Mais leurs exigences sont souvent contradictoires. L'analyse linguistique aboutit à des classements nuancés, qui comportent toujours des zones d'indétermination; la matière sur laquelle elle opère est éminemment continue, et il est rare qu'on puisse y tracer des limites nettes ; elle exige la plupart du temps un examen attentif de l'entourage syntagmatique [...] et paradigmatique [...] avant de trancher. La statistique, dans toutes ses applications, ne va pas sans une certaine simplification des catégories ; elle ne pourra entrer en action que quand le continu du langage a été rendu discontinu [...].

47.1 Normes de dépouillement

Malgré les connotations véhiculés par le mot *norme* dans le domaine linguistique, la notion de *norme de dépouillement* doit être ici comprise comme une exigence de standardisation provisoire des textes contenus dans un corpus. Cette standardisation est destinée avant tout à les rendre comparables, à les stabiliser le temps d'une expérience.

Nous allons illustrer sur un court extrait de ***Mitterrand1***, les problèmes liés à l'établissement d'une telle norme. Le premier fragment de texte (état A) correspond au texte tel qu'il a été saisi au départ.

> **État A :** Texte de départ
>
> Je crois qu'on ne peut que souhaiter cela. Le 14 juillet, c'est sans aucun doute - et c'est fort important - l'occasion d'une revue, d'un défilé, d'une relation directe entre notre armée et la nation.

Le second (Norme B) montre le même extrait du corpus après quelques transformations de surface destinées à permettre l'identification automatique des mêmes formes indépendamment de leur position dans la phrase (les majuscules de début de phrase ont été transformées en minuscules). Les barres verticales matérialisent la segmentation des unités.

> **Norme B :** Elimination des majuscules de début de phrase
>
> je | crois | qu'on | ne | peut | que | souhaiter | cela | . | le | 14 | juillet | , | c'est | sans | aucun | doute | - | et | c'est | fort | important | - | l'occasion | d | '| une | revue | , | d'| un | défilé, | d'| une | relation | directe | entre | notre | armée | et | la | nation.

Dans une phase suivante (Norme C), on a réuni certaines unités

polylexicales.

Norme C : Regroupement d'unités polylexicales

je | crois | qu' | on | ne | peut | que | souhaiter | cela | | . | le | 14 | juillet | , | c'est | sans aucun doute | - | et | c'est | fort | important | - | l' | occasion | d' | une | revue | , | d' | un | défilé | , | d' | une | relation | directe | entre | notre | armée | et | la | nation | . |

Dans les deux états suivants, les mots du texte ont été remplacés par des étiquettes (respectivement : des lemmes – Norme D – et des catégories grammaticales – Norme E).

Norme D : Lemmatisation

je | croire | que | on | ne | pouvoir | que | souhaiter | cela | . | le | quatorze | juillet | ce | être | sans | aucun | doute | - | et | ce | être | fort | important | - | le | occasion | de | un | revue | , | de | un | défilé, | de | un | relation | direct | entre | notre | armée | et | le | nation | .

Norme E : Catégorisation en parties du discours

{pronom} | {verbe} | {subordonnant} | {pronom} | {adverbe} | {verbe} | {subordonnant} | {verbe} | {pronom} | {ponctuation} | {déterminant} | {numéral} | {nom} | {pronom} | {verbe} | {préposition} | {déterminant} | {nom} | {ponctuation} | {coordonnant} | {pronom} | {verbe} | {adverbe} | {adjectif} | {ponctuation} | {déterminant} | {nom} | {préposition} | {déterminant} | {nom} | {ponctuation} | {préposition} | {déterminant} | {nom} | {ponctuation} | {préposition} | {déterminant} | {nom} | {adjectif} | {préposition} | {déterminant} | {nom} | {coordonnant} | {déterminant} | {nom} | {ponctuation}

Le dernier état du texte résulte d'un étiquetage permettant d'identifier les occurrences de quelques indices énonciatifs.

Norme F : Repérage d'indices énonciatifs

{embrayeur} {non-personne} {non-personne} {non-personne} {non-personne} {embrayeur}

Remarquons que, dans le cas de la mise en oeuvre de cette dernière norme de dépouillement, il ne s'agit plus d'une segmentation du texte de départ**.**

47.2 Décomptes automatisés

A la phase de délimitation des unités (qui peut être une segmentation) succède une phase de regroupement de celles que l'on considère comme identiques le temps de l'expérience (identification).

Pour un même texte, les différentes normes de dépouillement ne conduisent pas aux mêmes décomptes. Dans chaque expérience pratiquée, ces normes ne présentent pas le même degré de pertinence, ni les mêmes avantages (ou inconvénients) quant à leur mise en oeuvre. Néanmoins, au-delà des considérations propres à chaque domaine, une fois définie la norme de dépouillement et sa jurisprudence, les méthodes de la *statistique* s'appliquent de manière aveugle aux décomptes réalisés à partir de chacune des normes.

Comme on peut le voir sur les index réalisés à partir de ces transformations du texte de départ, le système des fréquences des unités soumises aux décomptes dépend étroitement de la norme de dépouillement retenue.

On voit sur ce petit exemple la grande latitude des choix possibles quand aux types de décomptes que l'on peut opérer à partir d'un même texte muni d'annotations. Pour chaque recherche particulière, ces choix résultent avant tout des objectifs de recherche poursuivis.

Norme A		**Norme B**		**Norme E**		**Norme F**	
,	4	,	4	{préposition}	15	{non-personne}	4
d’	3	d’	3	{déterminant}	8	{embrayeur}	2
c’	2	-	2	{nom}	8		
est	2	*c’_est*	2	{ponctuation}	6		
et	2	.	2	{pronom}	5		
une	2	et	2	{verbe}	5		
14	1	une	2	{adverbe}	2		
armée	1	nation	1	{coordonnant}	2		
aucun	1	ne	1	{subordonnant}	2		
cela	1	notre	1	{adjectif}	2		
.....		.....		{numéral}	1		
34 types		31 types		11 types		2 types	
45 occ.		40 occ.		56 occ.		6 occ.	

47.3 Incidence de la norme sur les décomptes

Mitterrand1 a été soumis à des dépouillements prenant en compte les différents systèmes d'unités évoqués plus haut. On a utilisé successivement :

- le système des caractères qui servent à encoder le texte sur support magnétique ;
- la segmentation du texte en formes graphiques obtenue en déterminant un ensemble de caractères délimiteurs (le point, la virgule, le point et virgule, etc.) ;
- la segmentation du texte en « lemmes » obtenue selon un ensemble de règles fixées par (Labbé, 1995) ;
- un système d'annotations grammaticales comportant 15 catégories différentes (nom, verbe, etc.) élaborée dans le cadre de cette même étude.

Le tableau 2 permet une comparaison rapide entre ces différents décomptes effectués à partir de niveaux d'annotation différents.

Tableau 2.— Décomptes sur Mitterrand1[277]

	caractères	formes	lemmes	catégories
nombre des occurrences :	1 667 251	297 258	307 865	307 865
nombre des types :	98	13 590	9 309	15
nombre des hapax[278] :	0	5 543	3 255	0
fréquence maximale :	224 865*	11 544	29 559	86 700 *

Les différents systèmes de décomptes produisent des descriptions difficilement comparables. Le système des catégories compte en effet un nombre relativement faible de types différents, les deux systèmes de descripteurs « lexicaux » (formes et lemmes) ont en commun de posséder un nombre très élevé de types s'étalant sur une large gamme de fréquence.

47.4 Exemple : l'accroissement du vocabulaire

Le problème de l'accroissement du vocabulaire (apparition de formes nouvelles au fur et à mesure que l'on avance dans la lecture du corpus) a été largement étudié dans les travaux de la statistique textuelle. La figure 2 rend compte de l'accroissement du vocabulaire, mesuré en lemmes et en formes graphiques. Les deux courbes ont la même *allure générale*. À un accroissement relativement fort au début du corpus, succèdent des périodes d'accroissement plus modestes, bien que tout allongement du corpus entraîne toujours l'apparition de nouvelles formes. Le nombre de formes nettement inférieur dans le cas du corpus lemmatisé fait que la deuxième courbe est toujours largement située en dessous de la première. En fait, deux tendances contraires influent sur les rapports qu'entretiennent ces nombres :

- le repérage de certaines unités composées de plusieurs formes graphiques (*à l'instar, à l'envi, d'abord, d'ailleurs*, etc.) tend à réduire le nombre des occurrences du corpus lemmatisé ;
- à l'inverse, l'éclatement en plusieurs unités distinctes de chacune des nombreuses occurrences des formes graphiques contractées (*au = à + le, des = de + les, etc.*.) tend pour sa part à augmenter le nombre des occurrences du corpus lemmatisé par rapport au texte initial.

[277] Les décomptes suivi de l'astérisque résultent d'une approximation statistique.
[278] Du grec *hapax legomenon* : chose dite une fois.

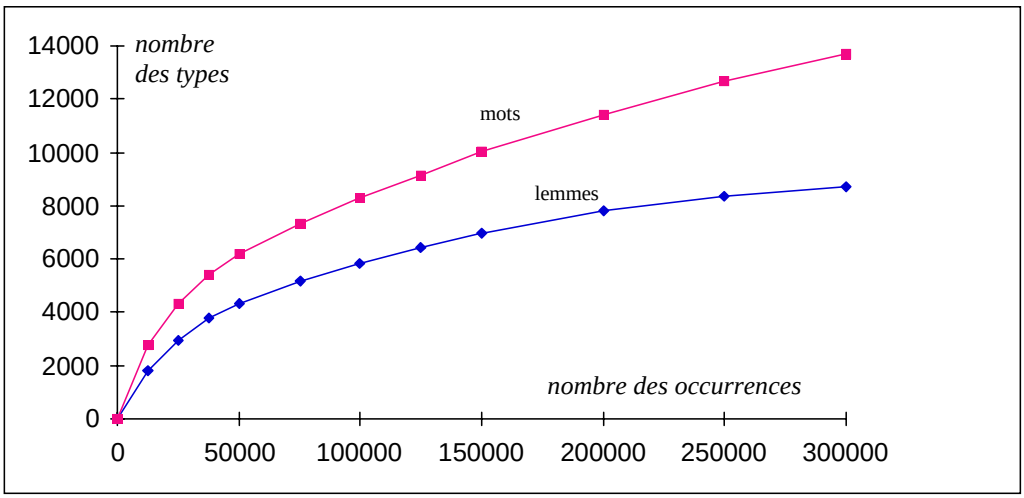


Figure 2. — L'accroissement du vocabulaire mesuré en formes graphiques et en lemmes

Cet exemple souligne la nécessité de pratiquer des comparaisons sur des comptages réalisés selon des normes de dépouillement identiques.

48. MESURES DE RECURRENCE SUR L'AXE SYNTAGMATIQUE

Les opérations de comptage des unités dans un corpus passent nécessairement par une phase de délimitation qui isole ces dernières de leur contexte immédiat. L'expérience montre cependant qu'après cette phase préliminaire, il est intéressant d'étudier en outre les récurrences et cooccurrences d'unités composées (suite de catégories syntaxiques, locutions ou expressions figées qui infléchissent, voire modifient totalement leurs significations) sous l'angle de leurs répétitions éventuelles dans le corpus.

48.1 Séquences d'unités

Au plan lexical, par exemple, les récurrences d'unités comme : *sécurité sociale, niveau de vie,* etc., sont dotées, dans les textes socio-politiques, d'un sens que l'on ne peut déduire à partir du sens des formes qui les composent.

On appelle *segment répété* toute suite d'unités textuelles reproduite sans variation à plusieurs endroits d'un corpus. Le nombre des unités qui composent le segment est sa *longueur*.

On peut recenser les segments répétés constitués par les unités qui relèvent de chacun des systèmes d'annotation dont on dispose sur le texte. Les suites de catégories grammaticales, par exemple, considérées sous l'angle de leur répétition dans le corpus renseignent sur la fréquence relative des constructions syntaxiques[279].

La recherche systématique des segments répétés de ***Mitterrand1***, parmi les formes lexicales, fait ainsi apparaître un très grand nombre de récurrences de fréquence élevée. Tous ces constats de répétition ne renvoient pas au même niveau d'analyse linguistique. Certains résultent de l'utilisation de syntagmes relativement bien formés, d'autres sont produits par la reprise partielle dans des phrases différentes de fragments plus ou moins autonomes au plan syntaxique.

Dans le tableau 3, on a rassemblé quelques-uns des segments qui sont à la fois longs et fréquents dans ce corpus. La colonne L donne la longueur du segment mesurée en formes graphiques, la colonne F indique sa fréquence.

Tableau 3. — Quelques segments fréquemment répétés dans **Mitterrand1**

L	F	segment
7	13	j ai dit tout à l heure
7	11	l ai dit tout à l heure
6	42	il n y a pas de
6	15	ce n est pas moi qui
6	15	je suis président de la république
6	15	que le président de la république
5	106	il n y a pas
5	93	le président de la république
5	36	dit tout à l heure
5	36	mais ce n est pas
5	34	de ce point de vue
4	366	ce n est pas
4	211	président de la république
4	190	je n ai pas
4	146	il n y a
4	124	un certain nombre de
4	121	tout à l heure

[279] On s'étonne par exemple, lors de l'analyse d'***Enfants,*** de ne pas trouver de segments répétés comprenant des verbes dans les réponses spécifiques (cf. *infra*) des plus diplômés

48.2 Quasi-segments

A côté des séquences reprises à l'identique à plusieurs endroits du corpus, on trouve des séquences qui sont l'objet de reprises partielles : la séquence je {catégorie=verbe} fermement que*,* par exemple, peut se réaliser sous la forme je pense fermement que, je crois fermement que*,* etc. Bécue (1993) a proposé un algorithme qui repère des *quasi-segments* (répétés). Cet algorithme permet, par exemple, de rassembler en une même unité (faire {lemme=<1>}$^{+}$ sport*)* les séquences comme faire du sport et faire un peu de sport*,* etc. Cependant, les quasi-segments sont encore plus nombreux que les segments, et leur recensement pose des problèmes de sélection et d'édition.

48.3 Cooccurrences

Pour une unité-pôle donnée, plusieurs méthodes permettent de sélectionner d'autres unités textuelles qui ont fortement tendance à se trouver dans un même voisinage que cette unité[280]. Le principe général de ces méthodes est le suivant. Pour sélectionner les formes cooccurrentes d'une forme-pôle, on commence par définir une unité de contexte, ou voisinage, à l'intérieur duquel on considérera que deux unités sont cooccurrentes. Cette unité de contexte peut correspondre à la phrase ou encore être constituée par un contexte de longueur fixe (*k* occurrences avant, et *k* occurrences après la forme-pôle). L'espace de cooccurrence peut également être défini de manière à ne pas dépasser les limites d'un constituant syntaxique. Si l'on se donne, à partir de l'exemple présenté plus haut (section 2.1), une fenêtre de deux occurrences avant et après la forme-pôle est (laquelle compte 2 occurrences), on construit autour de chacune des occurrences de la forme *est*, deux fenêtres matérialisées par les contextes compris entre les barres verticales :

```
Le 14             | juillet, c' est sans aucun | doute
sans aucun doute | - et c' est fort important | - l'
```

Dans ce cas, on sélectionne les cooccurrences de la forme-pôle avec les formes : juillet, c', sans, aucun, et , c', fort, important*.* Si l'on décide, toujours à partir de ce même extrait, de borner l'espace de cooccurrence au syntagme nominal minimal autour de la forme-pôle notre, on obtient une cooccurrence unique avec la forme armée.

Plusieurs méthodes statistiques se fixent pour but l'extraction des cooccurrences les plus remarquables dans un corpus de textes. Cette extraction s'appuie en général sur la comparaison des sous-ensembles de contextes qui contiennent l'unité-pôle avec ceux desquels elle est absente.

[280] Les applications de ces méthodes à l'étude de cooccurrences entre d'autres unités linguistiques devront faire l'objet d'études au cas par cas.

Pour chaque unité-pôle, on sélectionne ainsi un ensemble d'unités qui se trouvent situées de manière privilégiée dans les mêmes unités de contexte[281].

48.4 Filtrage des résultats

La sélection automatisée des segments répétés, quasi-segments et cooccurrences fréquemment attestés dans un corpus produit des listes d'unités qui renvoient en général à des niveaux très différents de l'analyse linguistique (lexies plus ou moins figées, tournures syntaxiques récurrentes, tournures de rhétorique etc.). Pour réduire le volume des listes ainsi constituées, certains chercheurs ont entrepris de constituer des procédures de filtrages applicable à ces listes afin d'en extraire, par exemple, les seuls éléments qui correspondent à des syntagmes bien formés :

~~ce n est pas moi qui~~

je suis président de la république

~~que le président de la république~~

49. COMPARER DES DECOMPTES AU SEIN D'UN CORPUS PARTITIONNE

Pour apprécier la répartition d'une unité linguistique à l'intérieur d'un corpus, il est nécessaire d'établir des comparaisons avec l'ensemble des unités de même type contenues dans le corpus. Une unité ne peut être jugée fréquente (ou rare) dans un texte que par comparaison avec d'autres unités dans ce même texte ou dans d'autres textes.

En pratique, ces comparaisons sont souvent malaisées du fait qu'il faut apprécier des décomptes qui concernent des unités dont les fréquences varient fortement dans des textes dont la longueur peut elle-même être très variable.

Le logiciel THIEF d'Étienne Brunet permet, par exemple, d'étudier la répartition de chacune des formes attestées dans le corpus du *Trésor de la Langue Française* parmi dix tranches chronologiques prédéfinies. On trouve figure 3 l'histogramme d'un indice qui permet de juger de la répartition de la

[281] Lafon (1984) et Labbé (1990) proposent des méthodes destinées à extraire les couples d'unités lexicales qui se rencontrent souvent à l'intérieur d'une même phrase. Church et Hanks (1990) utilisent, dans le même but, l'information mutuelle issue de la théorie de la communication de R. Shannon.

forme *gloire* dans ces dix tranches[282].

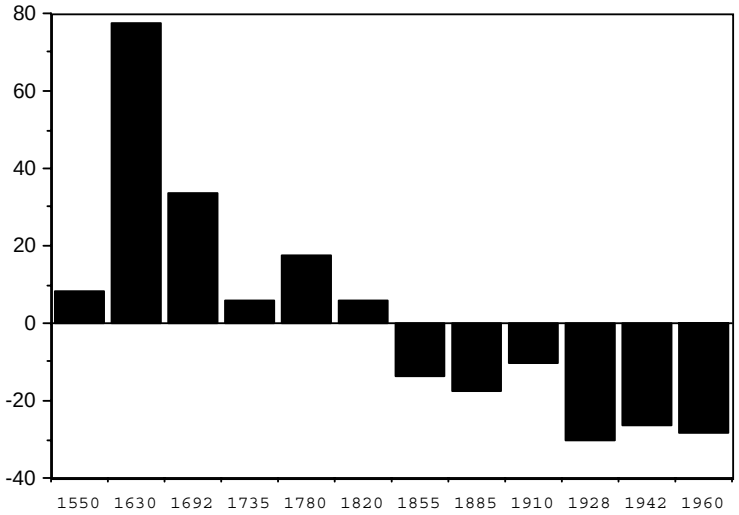


Figure 3. — La forme gloire *dans dix tranches du TLF*

Cette représentation graphique du phénomène appelle une interprétation très simple. La forme est tombée dans une désuétude relative au fil des périodes considérées.

La multiplication de résultats de ce type, à propos de formes différentes, incite à poser au corpus des questions plus générales. Quelles sont les formes qui subissent un sort similaire au cours des mêmes périodes ? Quelles sont celles qui au contraire voient le nombre de leurs occurrences augmenter relativement ?

Pour répondre de manière plus globale à des questions de ce type, il faut recourir aux méthodes de la statistique multidimensionnelle. Le point de départ des différentes méthodes qui servent à organiser la description comparative des parties d'un corpus est un tableau à double entrée que l'on constitue en croisant les parties du corpus et les différents types qui constituent le système d'unités préalablement choisi.

[282] Le calcul d'écart-réduit employé ici compare l'écart de la répartition observée dans chaque tranche à une répartition théorique.

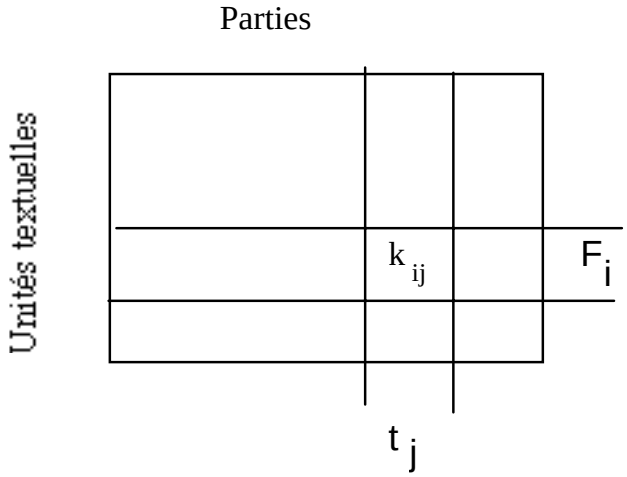


Figure 4. — Tableau de départ pour les analyses statistiques

A l'intersection de la ligne correspondant à l'unité *i* et de la colonne correspondant à la partie *j,* on trouve un nombre k_{ij} égal à la fréquence de l'unité *j* dans la partie *i* du corpus. La fréquence de l'unité *i* dans le corpus est égale à F_i. La longueur de la partie *j* (somme de toutes les occurrences de la partie j est égale à t_j.

49.1 Organiser la partition du corpus

A partir d'un même corpus, il est possible de constituer toute une série de partitions différentes (par émetteur ou par groupe d'émetteurs, si le corpus est plurilocuteur, en fonction de la date de rédaction, etc.). On peut ensuite décrire chacune des parties ainsi constituées par des systèmes de décomptes faisant intervenir des unités de différents niveaux (lemmes, formes graphiques, catégories grammaticales, ou tout autre type d'annotation). Le problème de la partition effective du corpus revêt une importance toute particulière dans la mesure où il s'agira ensuite d'étudier le contraste entre les parties découpées dans le corpus. La partition réalisée, on n'observera ensuite que des différences entre fragments du corpus ayant fait l'objet d'un même regroupement.

De son côté, la sélection d'un système d'unités linguistiques organise la comparaison des parties sur un plan d'analyse déterminé par les objectifs de la recherche. Les paragraphes qui suivent exposent brièvement les principes généraux du fonctionnement de ces méthodes sur des exemples empruntés à ***Enfants***.

En regroupant, par exemple, au sein d'une même partie les réponses fournies par les individus qui ont obtenu un diplôme équivalent, on réalise une

partition du corpus en trois parties (Aucun, Baccalauréat, Supérieur). Cette partition permet ensuite d'étudier les variations entre agrégats de réponses.

49.2 Repérer les faits saillants

La méthode des spécificités (Lafon, 1980) permet de mettre en évidence les cases du tableau de départ dont l'effectif est particulièrement élevé (spécificités positives) ainsi que celles dont l'effectif est au contraire anormalement faible (spécificités négatives). Elle s'applique successivement à chacune des cases du tableau décrit plus haut. Pour calculer le diagnostic relatif à l'effectif constaté pour une unité dans une partie donnée, on prend en compte la comparaison de quatre nombres :

- k_{ij} – sous-fréquence de l'unité dans la partie considérée.
- F_i – fréquence de l'unité dans l'ensemble du corpus.
- t_j – nombre des unités dans la partie
- T – nombre total des unités du corpus

Un calcul de type probabiliste permet de porter un jugement sur l'effectif contenu dans la case analysée (k_{ij}) compte tenu des trois autre nombres (F_i, t_j, T). Si l'effectif k_{ij} se situe dans les limites de ce que le calcul permettait d'espérer, on dit que la répartition constatée est *banale* (ce que l'on note « b »). Si ce n'est pas le cas, on calcule un indice de spécificité de la forme : +/-xx où :

- \+ indique une spécificité positive (sur-représentation par rapport à ce que les nombres (F_i, t_j, T) laissaient prévoir ;
- – indique une spécificité négative (sous-représentation) ;
- xx est un indice de spécificité qui est d'autant plus élevé que la sous-fréquence analysée s'écarte d'une répartition « neutre » qui est sous-jacente au modèle des spécificités[283].

Les constats de spécificités établis pour une même unité à propos de chacune des parties du corpus permettent de décrire le comportement de cette unité au sein du corpus. On voit ci-dessous les diagnostics de spécificités obtenus dans chacune des parties pour la forme *problèmes* qui compte 108 occurrences dans l'ensemble du texte.

	Aucun	Baccalauréat	Supérieur	Total
problèmes	41	20	47	108
diagnostic	-03	b	+04	
effectif (= t_j)	8006	3111	4487	15604

Ces résultats indiquent que la forme graphique *problèmes* est sous-

[283] Le modèle probabiliste utilisé pour juger de cette répartition est ici le modèle hypergéométrique, couramment utilisé dans ce type d'application.

représentée (-03) chez les sujets sans diplôme. Elle est au contraire sur-représentée (+04) chez les plus diplômés. La notation b en regard de la catégorie Baccalauréat indique que l'effectif des occurrences de *problèmes* dans cette catégorie n'est ni excessivement élevé ni excessivement bas. Nous verrons plus loin comment organiser entre eux les différents constats de ce types obtenus à partir de différents systèmes d'unités.

Tableau 4. — Formes spécifiques pour les répondants les plus diplômés

	F	f	Sp.
sur-emplois			
financières	174	79	+06
problèmes	108	47	+04
et	205	77	+03
face	10	8	+03
fait	25	14	+03
couple	95	39	+03
raisons	178	66	+03
affective	12	8	+03
difficultés	83	37	+03
responsabilités	22	13	+03
sous-emplois			
vie	180	35	-03
NON-REP	65	10	-03
le	474	111	-03
n	94	16	-03
vois	20	0	-03
manque	160	29	-03
aucune	33	3	-03
sais	25	1	-03
y	57	7	-03
faire	22	1	-03
pas	325	71	-03
emploi	79	13	-03
a	74	12	-03
travail	152	26	-04
il	105	15	-04
chômage	285	52	-05

Une fois ce calcul effectué pour chacune des cases du tableau analysé, le regroupement des diagnostics relatifs à une même partie fournit une description de cette partie par la mise en évidence des termes qu'elle sur-emploie, ainsi que celle des termes qu'elle sous-emploie[284]. Voici, à titre d'exemple, dans le tableau 4 ci-dessous, les formes jugées spécifiques, c'est-à-dire les formes tout particulièrement sur-représentées (resp. sous-

[284] On trouve un panorama des applications de ces méthodes aux textes socio-politiques dans (Habert, 1985).

représentées) dans la partie du corpus qui correspond aux plus diplômés.

50. APPROCHES MULTIDIMENSIONNELLES

Chacune des dimensions du tableau rectangulaire considéré plus haut permet de définir des distances (ou des proximités) entre les éléments de l'autre dimension[285]. Ainsi, l'ensemble des colonnes (dans notre cas les parties du corpus) permet de définir à l'aide de formules appropriées des distances entre lignes (ici les unités appartenant à un système d'annotation). De la même façon, l'ensemble des lignes permet de calculer des distances entre colonnes.

On obtient ainsi des tableaux de distances, auxquels sont associées des représentations géométriques complexes décrivant les similitudes existant entre les lignes et entre les colonnes des tableaux rectangulaires à analyser.

Le problème est alors de rendre assimilables et accessibles à l'intuition ces représentations, au prix d'une perte de l'information de base qui doit rester la plus petite possible.

Deux familles de méthodes permettent d'effectuer ces réductions :

- *Les méthodes factorielles* produisent des représentations graphiques sur lesquelles les proximités entre points-lignes et entre points-colonnes traduisent les associations statistiques entre lignes et entre colonnes ;
- *Les méthodes de classification* opèrent des regroupements en classes (ou en familles de classes hiérarchisées) des lignes ou des colonnes.

50.1 Classer les unités et les textes

Les méthodes de classification ascendante hiérarchique s'appliquent aux tableaux à double entrée décrits plus haut. On peut soumettre à la classification soit l'ensemble des colonnes du tableau (qui correspondent la plupart du temps aux différentes parties d'un corpus) soit celui des lignes de ce même tableau (lesquelles correspondent en général à un système d'unités textuelles recensées dans le corpus).

[285] En analyse des données, on utilise souvent une distance qui est une somme de carrés pondérés dite *distance du chi-deux*. Cette distance possède toute une série de propriétés particulièrement intéressantes (Lebart et Salem, 1994, p. 87).

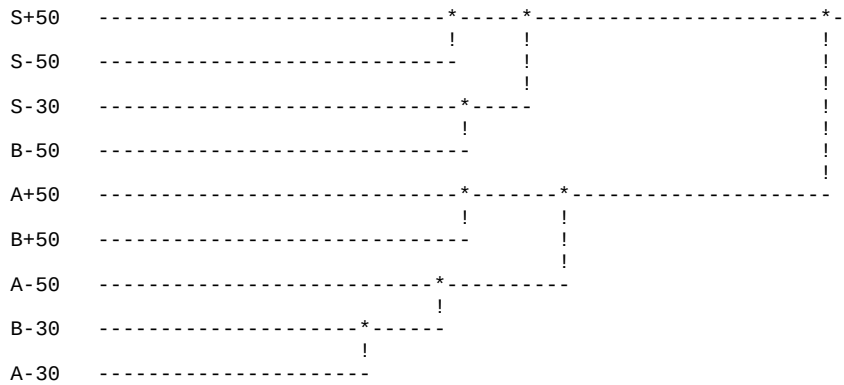
50.1.1 Classification ascendante hiérarchique

Dans le cas de la *classification ascendante hiérarchique*, on part d'un ensemble de *n* éléments, affectés chacun d'un poids proportionnel à leur importance dans l'ensemble, et entre lesquels on a calculé des distances. On commence par agréger les deux éléments les plus proches. Ce couple constitue alors un nouvel élément dont on peut recalculer à la fois le poids et les distances par rapport chacun des éléments qu'il reste à classer[286]. À l'issue de cette étape, le problème se trouve ramené à celui de la classification de *n-1* éléments. On agrège à nouveau les deux éléments les plus proches, et l'on réitère ce processus (*n-1* fois au total) jusqu'à épuisement de l'ensemble des éléments.

Chacun des regroupements effectués en suivant cette méthode s'appelle un *noeud*. L'ensemble des éléments terminaux rassemblés dans un noeud est une *classe*. La représentation de la classification sous forme d'arbre hiérarchique ou *dendrogramme* est la représentation la plus courante. L'interprétation d'une telle hiérarchie s'appuie sur l'analyse des seules distances entre éléments ou classes faisant l'objet d'un même noeud (*i.e*. seules les proximités entre éléments appartenant à une même classe peuvent être interprétées).

Appliquée au tableau analysé ci-dessus, la classification ascendante hiérarchique produit un regroupement en deux sous ensembles relativement distincts : les diplômés du supérieur d'une part et les sans-diplômes d'autre part. Les groupes de diplômes intermédiaires se répartissant entre ces deux sous-ensembles.

*Tableau 5 — Classification sur les parties d'***Enfants**



Les classifications effectuées sur l'ensemble des parties et celles réalisées à partir de l'ensemble des unités, répondent à des besoins d'analyse distincts qui entraînent, dans les deux cas, des utilisations différentes de la méthode.

[286]Dans la pratique il existe un grand nombre de façons de procéder qui correspondent à cette définition, ce qui explique la grande variété des méthodes de classification automatique, sur ces méthodes on peut consulter (Saporta, 1990, p. 241-261).

50.1.2 Classifications de formes

Lorsqu'il s'agit d'étudier des textes (littéraires, politiques, historiques), les classifications portant sur les formes d'un corpus concernent en général des ensembles dont la dimension dépasse très largement celle de l'ensemble des parties. L'arbre de classification réalisé à partir d'un tel ensemble se présente sous une forme relativement volumineuse qui complique considérablement toute synthèse globale. Dans la pratique, on abordera l'étude des classifications ainsi réalisées en considérant par priorité les associations qui se réalisent aux deux extrémités du dendrogramme :

- les classes du niveau inférieur de la hiérarchie constituées par des agrégations de formes agrégées dès le début de la classification et qui correspondent souvent à des associations de type cooccurrentielles ;
- les classes supérieures, souvent constituées de nombreuses formes, que l'on étudiera globalement.

Les associations réalisées aux premiers niveaux de la classification regroupent, par construction, des ensembles de formes dont les profils de répartition sont très similaires (proportionnels et parfois mêmes identiques) dans les parties du corpus. Le retour systématique au contexte permet seul de distinguer parmi ces associations celles qui proviennent essentiellement de la reprise de segments plus ou moins longs, celles qui sont générées par les cooccurrences répétées de plusieurs formes à l'intérieur de mêmes phrases ou de mêmes paragraphes et les associations qui résultent de l'identité plus ou moins fortuite de la ventilation de certaines formes.

La figure 6 montre une petite partie de l'arbre de classification réalisé à partir des formes les plus fréquentes dans ***Enfants.*** L'analyse du contenu de ces classes se fait en retournant fréquemment au contexte.

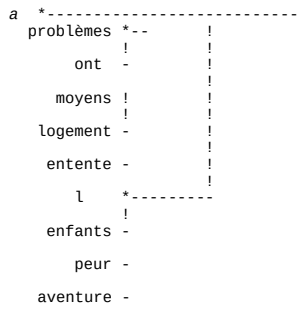


*Figure 6. — Extrait d'une classification sur les formes d'***Enfants**

50.1.3 Classifications descendantes

Certains auteurs (Reinert, 1990) utilisent d'autres procédures de classification pour analyser les corpus textuels. Le principe général de la méthode est le suivant. On commence par découper dans le texte des *unités de contexte* (la plupart du temps, une fenêtre comportant quelques occurrences à gauche et à droite de chaque occurrence du texte). L'ensemble de ces unités est ensuite divisé successivement en classes (de manière dichotomique à chaque étape). Ce processus aboutit à rassembler des formes qui ont tendance à se retrouver dans des contextes proches.

*50.2 **L'approche factorielle***

L'analyse factorielle des correspondances crée une typologie qui porte à la fois sur l'ensemble des parties du corpus et sur l'ensemble des unités par lequel ce dernier est décrit[287]. Négligeant toute une partie de l'information contenue dans le tableau des distances, cette méthode fournit des représentations approchées des distances calculées entre les éléments de chacun des deux ensembles mis en correspondance. Les graphiques-plans qui sont un des résultats fournis par l'analyse sont en quelque sorte les meilleures représentations bidimensionnelles possibles de chacun des ensembles. Sur ces graphiques, deux parties sont proches si elles emploient les mêmes unités dans des proportions semblables.

Cette méthode permet de créer une typologie qui peut s'affiner au fur et à mesure de la prise en compte des axes factoriels successifs. Elle est particulièrement adaptée à la mise en évidence des principales oppositions qui sous-tendent le corpus.

Remarquons que la classification ascendante hiérarchique et l'analyse factorielle sont des méthodes très complémentaires dans la mesure où l'une permet au chercheur de concentrer son attention sur les proximités locales pouvant exister entre chaque élément alors que la seconde rend compte des grandes oppositions pouvant exister dans le corpus.

Ainsi, les réponses contenues dans ***Enfants*** ont été regroupées cette fois en neuf parties qui correspondent au croisement de trois catégories de diplôme (A=aucun, B=Baccalauréat ou BEPC et S=Supérieur) avec trois catégories d'âge (moins de 30 ans, 30 à 50 ans, 50 ans et plus). On a ensuite calculé le tableau qui croise ces neuf catégories avec les formes du corpus[288].

---

[287] L'ouvrage de référence est le livre de J.-P. Benzécri et coll. (Benzécri, 1973). On trouvera des présentations différentes de cette même méthode destinées au lecteur non-mathématicien dans (Salem, 1987) ainsi que dans (Lebart et Salem, 1994).

[288] Pour alléger les résultats, seules les formes de fréquence supérieure à 10 occurrences ont été retenues. L'expérience montre que ce type de sélection a peu d'influence sur les résultats de l'analyse.

Commençons par un exemple très simple. On a représenté (Figure 7) les neuf parties du corpus en fonction de leur utilisation des formes : *raisons* (axe vertical) et *problèmes* (axe horizontal). La valeur portée sur chacun des axes est égale à la proportion d'utilisation (exprimée en 10 000èmes) de chacune de ces formes par chacune des parties. On voit que les parties ne se répartissent pas sur l'ensemble du graphique mais sont plutôt regroupées autour d'une des diagonales. Cela veut dire que l'emploi des deux formes par les émetteurs manifeste une *corrélation*. Ceux qui emploient beaucoup l'une des formes (S-30, S-50, c'est-à-dire les diplômés les plus jeunes) ont tendance à utiliser également l'autre (et inversement).

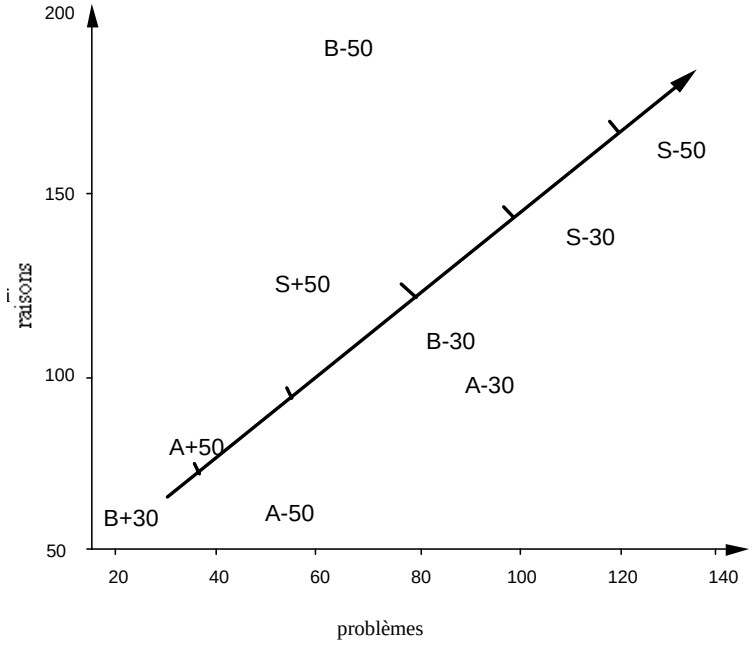


*Figure 7. – Les parties d'*Enfants *et les formes* raisons *et* problèmes*.*

Si l'on accepte de perdre un peu de l'information contenue sur ce graphique, on peut simplifier la représentation des parties en traçant un axe qui épouse *le mieux possible* la forme du nuage de points représenté sur la figure 7. Si l'on munit cet axe d'un système de coordonnées, on obtient une représentation des distances entre les parties (figure 8) qui est moins précise mais plus *synthétique*.

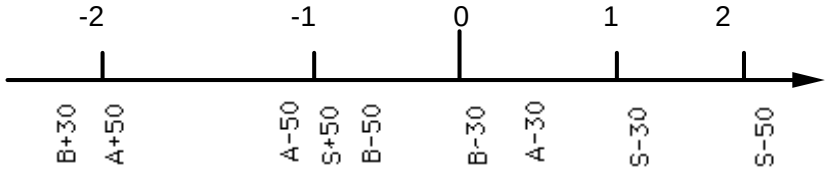


Figure 8. — Les mêmes parties disposées sur un « facteur »

Les méthodes factorielles opèrent, à partir des immenses tableaux soumis à l'analyse, des synthèses du même type. Partant d'un tableau qui compte cette fois plusieurs milliers de formes et toujours neuf parties, l'analyse des correspondances extrait une information synthétique. La représentation simplifiée des distances entre catégories met en évidence la principale information contenue dans le tableau de données soumises à l'analyse : la proximité (basée sur un usage proche du stock des formes lexicales) des agrégats proches par le diplôme ou par l'âge (figure 9).

Il faut comprendre que la méthode de calcul ne s'appuie à aucun moment sur des données extérieures lui permettant d'inférer des proximités entre tel ou tel agrégat. Les rapprochements sont effectués uniquement à partir des comparaisons du stock de vocabulaire employé par les répondants appartenant à un même agrégat âge / diplôme.

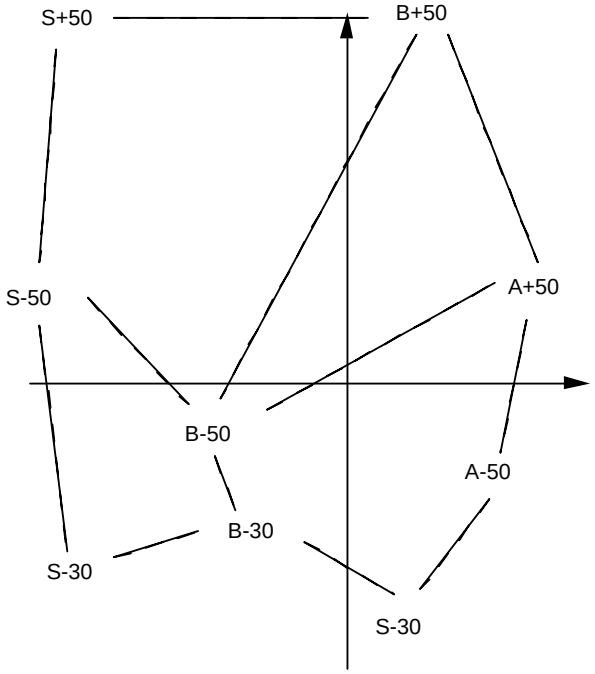


Figure 9. — Les 9 classes Age x Diplôme sur le plan des deux premiers facteurs de l'analyse.

Une représentation simultanée des formes et des parties sur le même graphique peut permettre de mettre en évidence les formes qui sont principalement responsables de cette typologie.

51. ARTICULER DES CONSTATS SUR DES UNITES DIFFERENTES

L'articulation des résultats obtenus à l'aide de telles méthodes à partir de différentes normes de dépouillement permet une description beaucoup plus sûre des contrastes entre les parties du corpus[289]. La typologie réalisée sur les parties dépend peu, dans le cas qui nous préoccupe, des variations dans la norme de dépouillement (lemme / formes graphiques, etc.). Loin de constituer une gêne pour l'interprétation, les éclairages complémentaires projetés par différent systèmes d'unités nous aident à mieux comprendre les oppositions pouvant exister entre les textes que l'on compare.

51.1 Articuler unités isolées et séquences d'unités

L'exemple qui suit montre comment articuler de tels décomptes dans le cadre de la méthode des spécificités, la plus simple des méthodes exposées jusqu'ici.

Les occurrences du segment répétés problèmes financiers peuvent être considérées comme un sous-ensemble des occurrences de la forme problèmes pour lesquelles une occurrence de la forme financiers apparaît immédiatement après. On peut appliquer au segment répété problème financiers le calcul des spécificités.

Pour les deux formes et le segment évoqués, ce calcul donne :

Forme / diplôme	Aucun	BACC	Sup.	F
problèmes	41 -03	20 b	47 +04	108
financiers	37 b	19 b	30 b	86
problèmes financiers	17 -03	11 b	23 +03	51

Comme on le voit, les diagnostics ci-dessus ne coïncident pas tous entre eux. Ils rendent compte de la diversité des associations réalisées dans le corpus. La forme financiers, par exemple, est considérée comme régulièrement répartie alors que le segment problèmes financiers et la forme problèmes sont plutôt sur-représentés chez les plus diplômés.

Le tableau 6 interclasse d'après un indice de spécificité calculé selon les mêmes procédures des diagnostics obtenus sur des formes et sur des segments répétés dans le corpus. L'avantage de ce second tableau sur son homologue réalisé à partir des formes simples est qu'il constitue un pas, réalisé automatiquement, vers la remise en contexte des résultats.

***Tableau 6**. — Formes et segments les plus caractéristiques pour les répondants les*

[289] Des résultats tout à fait similaires ont été obtenus dans une expérience du même type portant cette fois sur des décomptes de lemmes au sein de la même partition du corpus.

plus diplômés

	F	f	Sp.
financières	174	79	+06
les difficultés financières	19	14	+05
difficultés financières	32	19	+04
problèmes	108	47	+04
fait de	10	7	+03
et	205	77	+03
face	10	8	+03
et les	17	10	+03
du couple	48	23	+03
fait	25	14	+03
situation économique	24	13	+03
raisons financières	93	38	+03
couple	95	39	+03
raisons	178	66	+03
problèmes financiers	51	23	+03
affective	12	8	+03
les problèmes	35	18	+03
difficultés	83	37	+03
des responsabilités	13	9	+03
responsabilités	22	13	+03
le fait	16	11	+03

Ce tableau présente de nombreuses redondances qui résultent du fait que, dans un premier temps, les listes d'unités spécifiques sont produites de manière entièrement automatique, sans aucun filtrage. L'illustration par les segments répétés précise la signification des unités mises en évidence par le calcul des spécificités. L'implication des dénombrements portant sur les segments répétés permet d'extraire de l'enchevêtrement inextricable des segments répétés des unités qui précisent la description par les unités effectuée à partir des unités isolées de leur contexte immédiat.

51.2 Articuler différents systèmes d'unités

La comparaison entre les différentes parties d'un corpus devient encore plus lisible si l'on implique les décomptes réalisées pour chacune d'elles à l'intérieur de différents systèmes d'unités linguistiques[290].

De la même manière que nous l'avons fait ci-dessus, il est possible de compléter la description des parties du corpus par des comptages réalisés sur l'ensemble des annotations disponibles dans le corpus considéré. Le tableau 7 montre les mêmes opérations de sélection d'unités caractéristiques

[290] Cf. (Salem, 1987 ; 1993) et (Habert et Salem, 1995)

réalisées cette fois à partir des annotations de type grammatical et des segments constitués à partir de ces dernières.

Tableau 7. — Formes graphiques, lemmes, catégories grammaticales et segments répétés les plus caractéristiques pour les répondants les plus diplômés

	unités	F	f	Ind.
C	{nom} {adjectif}	863	312	+07
F	financières	174	79	+06
L	*financier virgule*	123	59	+06
F	les difficultés financières	19	14	+05
C	{nom} {adjectif} {ponctuation}	32	20	+05
L	*le difficulté financier*	19	14	+05
F	problèmes	108	47	+04
F	difficultés financières	32	19	+04
C	{adjectif} {coord} {adjectif}	20	13	+04
C	{coord} {adjectif}	26	16	+04
C	{nom} {adjectif}{coord} {adjectif}	19	13	+04
C	{determinant ind} {nom} {adjectif}	36	20	+04
L	*difficulte financier virgule*	12	10	+04
L	*que ce*	26	17	+04
L	*difficulte financier*	32	19	+04
L	*financier*	374	136	+04
L	*probleme*	145	60	+04
F	problèmes financiers	51	23	+03
F	couple	95	39	+03
F	responsabilités	22	13	+03
F	raisons financières	93	38	+03
F	situation économique	24	13	+03
F	affective	12	8	+03
F	du couple	48	23	+03
F	et	205	77	+03
F	monde	16	10	+03
F	des responsabilités	13	9	+03
F	difficultés	83	37	+03
F	les problèmes	35	18	+03
F	et les	17	10	+03

Légende : La colonne de gauche indique la nature des unités et séquences d'unités prises en compte selon le code suivant : F – formes graphiques, L – lemmes, C – catégories grammaticales.

Comme plus haut, les unités sélectionnées dans ce tableau l'ont été en raison de leur abondance particulière dans la partie du corpus qui correspond aux plus diplômés. L'interclassement des unités selon l'indice de spécificité calculé de la même manière sur tous les types d'annotations et sur les segments réalisés à partir de ces dernières permet de classer l'ensemble des constats du plus surprenant au plus banal.

La redondance s'est encore accrue mais la description est devenue plus beaucoup plus riche, faisant intervenir de plusieurs niveaux de l'analyse linguistique.

52. **TEMPS LEXICAL**

Certains corpus réunis par échantillonnage au cours du temps d'une même source textuelle présentent dès le départ une homogénéité remarquable : les textes réunis sont produits dans des conditions d'énonciation très proches, parfois par le même locuteur. Leur étalement dans le temps doit permettre de mettre en évidence ce qui varie au cours du temps. Nous appelons ces corpus des *séries textuelles chronologiques*. ***Mitterrand1*** constitue, nous l'avons vu, un corpus de ce type.

Dans le cas des telles séries, les résultats factoriels font apparaître un schéma d'évolution chronologique qui rend compte de l'existence d'une évolution. Les apparitions, disparitions ou fluctuations des formes s'effectuent de manière suffisamment organisée, au regard du temps, pour que les périodes consécutives apparaissent plus proches dans l'emploi qu'elles font du vocabulaire que les périodes séparées par un intervalle de temps plus long.

La figure 10 montre des résultats issus d'une AFC portant sur les formes de fréquence supérieure ou égale à 5 occurrences dans ***Mitterrand1***. On le voit, les périodes consécutives sont plutôt proches les unes des autres. L’ensemble des points dessine une ligne incurvée en son centre.

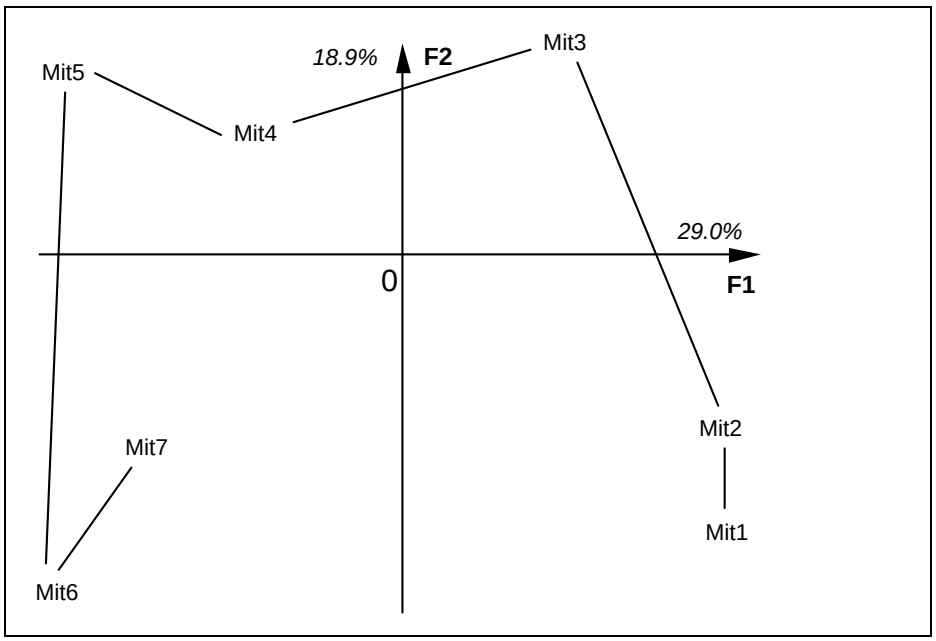


Figure 10. — Les deux premiers facteurs issus de l'analyse des correspondances[291]

Pour avancer dans l'analyse, il faut créer des procédures permettant d'exhiber les unités textuelles responsables de cette évolution d'ensemble.

52.1.1 Accroissements spécifiques

Le calcul des accroissements spécifiques permet de repérer les changements brusques dans l'utilisation d'un terme lors d'une période donnée par rapport à l'ensemble des périodes qui précèdent. Pour chaque terme dont la fréquence dépasse un seuil fixé à l'avance, pour chaque période du corpus à partir de la seconde, on compare, selon le modèle des spécificités présenté plus haut, la sous-fréquence observée dans la période considérée à la fréquence de cette même unité dans l'ensemble des périodes précédentes.

Le tableau 8 donne quelques accroissements spécifiques majeurs pour l'ensemble de ***Mitterrand1***. Les accroissements spécifiques sont notés à l'aide des symboles : / et \ qui indiquent des spécificités respectivement positive et négative de l'accroissement ; (*i.e*. un sur-emploi et un sous-emploi spécifique par rapport aux parties précédentes). La dernière colonne indique la période (`pér.`) *i.e.* la partie du corpus concernée par le diagnostic d'accroissement spécifique. Pour chaque terme, la colonne Fx donne le nombre des occurrences de ce terme dans le groupe de périodes

[291] Il s'agit des deux premiers facteurs issus de l'analyse du tableau croisant formes graphiques de fréquence supérieure à 20 et périodes (1 397 formes x 7 périodes).

précédentes.

Tableau 8. — Chronique des spécificités maximales pour **Mitterrand1**

terme	F	Fx	f	spec.	pér.
nationalisations	42	31	0	/12	2
israel	71	56	2	\11	3
monsieur	430	213	91	/11	4
nouvelle calédonie	33	22	20	/11	4
référendum	27	19	18	/11	4
très	627	329	127	/11	4
chaîne	39	36	34	/19	5
la france	1016	722	106	\11	5
la majorité	91	70	45	/12	5
notre	442	337	35	\11	5
nous	2059	1700	308	\11	5
avons	523	488	30	\11	6
étudiants	28	28	27	/21	6
majorité	212	149	90	/20	5
nous	2059	1877	177	\17	6
oeuvres	29	24	19	/11	6
pour 100	204	195	2	\12	6
arabe	34	34	23	/13	7
l iran	50	50	41	/27	7
monde arabe	21	21	17	/12	7
nous	2059	2059	182	\12	7

Pour une période donnée, la liste des accroissements spécifiques de la période renseigne sur l'émergence d'un vocabulaire particulier. Le tableau 9 donne les accroissements ainsi calculés pour la 7e partie du corpus constituée par des interventions effectuées au cours des années 1987-1988.

Tableau 9. — Accroissements spécifiques majeurs pour la 7e période de **Mitterrand1**

```
       l iran                  50       41   /27
       iran                    53       41   /25
       arabe                   34       23   /13
       monde arabe             21       17   /12
       d instruction           20       16   /11
       instruction             23       17   /11
       l irak                  29       18   /09
       irak                    32       18   /08
       élection                35       18   /07
       président              303       73   /07
       d armes                 27       15   /07
       un président            28       15   /07
       politiques             105       34   /07
       armes                   93       32   /07
       juge                    35       17   /07
       pays                   748      151   /07
-------------------------------------------------
       nous avons             413       27   \06
       inflation               83        0   \06
       avons                  523       35   \07
       jeunes                 134        2   \07
       nous                  2059      182   \12
```

52.1.2 Formes chrono-homogènes

Les méthodes présentées ci-dessus permettent de décrire, au fil des périodes, l'évolution des unités textuelles que l'on peut recenser dans un corpus chronologique. Les schémas d'évolution établis pour chacune des unités font apparaître des ensembles d'unités qui ont tendance à évoluer de conserve au fil des périodes : les formes *chrono-homogènes*.

En fait, l'idée qui sous-tend cette approche est la suivante : pour des formes fréquentes dans le corpus, le fait que plusieurs formes évoluent de manière proportionnelle tout au long des périodes ne peut être mis au compte du hasard. Il faut donc, dans chaque cas, déterminer la cause profonde qui est à l'origine de ces regroupements. Selon les cas, on trouvera des groupements liés à une thématique, à une actualité, etc.

La figure 11 présente un groupe de formes, parmi les plus fréquentes de ***Mitterrand1***, qui sont chrono-homogènes par rapport à la forme je. On retrouve ici un ensemble de marqueurs de la première personne.

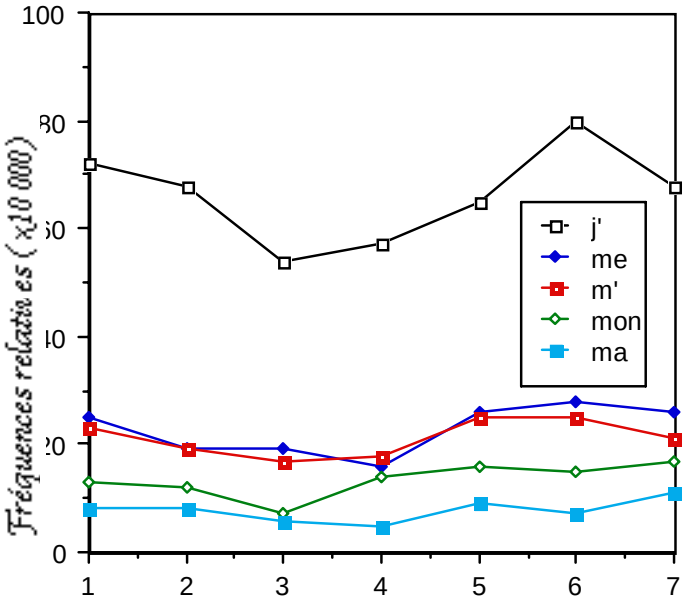


Figure 11. — Formes chrono-homogènes à la forme je *dans* **Mitterrand1**

L'étude des séries textuelles chronologiques s'opère donc en combinant plusieurs types de méthodes. L'analyse des correspondances permet de vérifier que le corpus chronologique, compte tenu d'une périodisation donnée, relève bien du schéma général d'évolution du vocabulaire. Elle permet également de localiser des écarts éventuels avec le schéma général, qui seront dans la plupart des cas sources d'interrogations utiles. L'examen attentif des accroissements spécifiques signale à la fois des moments particuliers dans l'évolution du vocabulaire et les unités textuelles qui en sont à l'origine. Enfin, l'étude des termes chrono-homogènes permet de constituer des classes d'unités et d'étudier leur évolution conjointe au fil des périodes.

53. CONCLUSION

Les analyses portant sur des textes annotés apportent un complément d'information important, par rapport aux mêmes analyses effectuées à partir d'un découpage en formes graphiques, dès lors qu'il s'agit de mettre en

évidence des unités textuelles caractéristiques pour chacune des parties d'un corpus de textes, encore que ces résultats soient difficiles à manier simultanément.

L'utilisation de comptages portant sur les segments répétés d'un corpus pour illustrer les typologies réalisées à partir des formes permet de dépasser les résultats obtenus sur les formes isolées de leur contexte immédiat et d'accéder à la description d'associations remarquables par leur répartition. Les différentes méthodes de calcul des cooccurrences concourent également à ce but.

Par exemple, dans le domaine de l'étude des textes politiques, l'expérience a montré que le singulier et le pluriel de certains substantifs renvoient souvent à des oppositions profondes au plan de l'idéologie politique. On peut dire que de grandes oppositions idéologiques se sont souvent exprimées à travers l'emploi du singulier ou du pluriel d'une même forme de vocabulaire. *Les classes ouvrières*, proclamait le pouvoir monarchique sous Louis-Philippe (1830-1848) ; *la classe ouvrière,* contestaient les organisations ouvrières. De même les années 1970 ont vu s'opposer les défenseurs *des libertés républicaines* (la gauche et les syndicats) aux défenseurs de *la liberté* avec, bien entendu, des contenus partiellement différents. Cette distinction est en revanche moins pertinente dans le cas de l'étude de ***Menelas*** : le comportement du singulier et du pluriel de *sténose* ne justifie pas qu'on les considère séparément.

L'éclairage qu'apporte l'approche quantitative à la connaissance d'un corpus de textes réunis à des fins de comparaison s'exprime de manière privilégiée sous forme de contrastes entre les unités que l'on peut décompter dans les parties du corpus. Ces circonstances fournissent indirectement un critère quant au choix des unités à retenir dans les analyses textuelles : si les différentes réalisations d'une unité linguistique sont distribuées de la même manière parmi les parties du corpus que l'on compare, il ne sert à rien de les distinguer dans les comptages, car elles ne seront pas à l'origine des contrastes mis en lumière par les analyses statistiques. Si par contre les réalisattions d'une même unité ont des ventilations très différentes à l'intérieur du corpus considéré, le fait de les réunir en une même unité statistique prive le chercheur de constats qui auraient pu l'intéresser.

CONCLUSION

G. Leech (1991, p. 25) souligne le tournant des années actuelles : « Ceux qui travaillent sur corpus électroniques se trouvent soudain dans un univers en pleine expansion. Pendant des années, la linguistique de corpus a été l'obsession d'un petit groupe qui recevait peu de soutien, que ce soit de la linguistique ou de l'informatique. » Ce constat vaut au tout premier chef pour le monde anglo-saxon. Mais si l'on fait le bilan du domaine couvert par les linguistiques de corpus, quelles perspectives s'ouvrent, en particulier pour la francophonie ?

54. BILAN

Face à un domaine riche en travaux d'horizons théoriques et méthodologiques variés – en TALN et en linguistique, nous ne prétendons pas avoir rendu compte des recherches les plus représentatives. Comment, face à un champ en pleine mouvance, en identifier les grandes tendances ? Il aurait fallu un recul dont nous ne disposons pas et qu'à notre avis, on ne peut pas encore prendre. Nous avons plutôt cherché à fournir une typologie de travaux prometteurs. Espérons que cette typologie puisse aussi servir de grille de lecture pour situer d'autres recherches que celles qui ont été directement évoquées.

54.1 Avancées

La robustesse est le maître mot des techniques d'annotation qui sont visées pour les textes tout-venant. On est loin de pouvoir en donner une définition précise. Néanmoins, l'examen des outils disponibles et des corpus annotés le montre : l'étiquetage est relativement bien maîtrisé actuellement, le parsage fruste progresse, même si les tâtonnements dominent encore pour les

traitements sémantiques.

Constatons que certaines tâches d'annotation sont progressivement automatisées, avec éventuellement des phases de pré- ou de post-traitement. On commence à mieux cerner ce qui est effectivement automatisable et ce qui ne le sera probablement jamais. C'est ce que nous avons vu avec l'acquisition terminologique (chapitre II) : la frontière entre le repérage automatique et ce qui relève de compétences humaines peu formalisables se précise.

Il est frappant de constater que certaines de ces avancées reposent sur des techniques somme toute relativement simples. On est étonné par l'écart entre les méthodes utilisées, parfois frustes, et la richesse des résultats, comme l'indique E. Brill (1995, p. 544) : « Les méthodes basées sur les corpus sont souvent capables de réussir tout en ignorant la complexité réelle du langage, en s'appuyant sur le fait que des phénomènes linguistiques complexes peuvent souvent être observés indirectement par le biais de simples épiphénomènes. » C'est le cas pour l'alignement de textes, qui utilise parfois « une corrélation très forte entre la longueur des segments qui sont mis en correspondance traductionnelle » (Isabelle et Amstrong, 1993), que cette longueur soit mesurée en nombre de mots ou en caractères. C'est le cas encore de la production d'ébauches d'entrées de dictionnaires par des méthodes comme celles utilisées par Grefenstette (1994).

Un autre point positif est le recul des illusions en ce qui concerne le traitement automatique de textes tout-venant. Les conditions instutionnelles à réunir, les performances des outils existants ainsi que le coût de l'obtention de corpus annotés sont désormais mieux connus. Les opérations d'évaluation des outils et des ressources qui ont été lancées dans le monde anglo-saxon et qui débutent pour la francophonie (Paroubek *et al.*, 1997) sont salutaires : elles fournissent des états de l'art sectoriels et précis.

L'observation raisonnée de données volumineuses enrichit la pratique linguistique. Elle fournit des données que l'intuition du linguiste aurait refusées (taxées d'inacceptables) ou qu'elle n'aurait pas prévues (variation d'expressions toutes faites et de termes). Elle accroît la précision des descriptions ou les rectifie (en linguistique diachronique par exemple). Elle rend manifeste le poids des différentes règles. Les traitements multidimensionnels permettent de repérer des corrélations inattendues et en tout cas non perceptibles directement entre des phénomènes langagiers relevant de niveaux distincts de l'analyse linguistique.

54.2 Limites

Les ressources pour le français sont encore denrée rare. Il n'existe pas d'équivalents pour le français de ***Brown***, ***LOB*** et de ***BNC***, pour la langue contemporaine, ou d'***Archer***, pour l'histoire de la langue, c'est-à-dire des corpus diversifiés, associant des registres différents et offrant aux linguistes

comme aux informaticiens des objets d'étude variés. Il n'existe pas non plus d'étiqueteur-lemmatiseur immédiatement accessible ni d'équivalent français de ***WordNet*** pour l'annotation sémantique. Le risque est que soient baptisés du nom de corpus des rassemblements de textes électroniques disponibles n'offrant pas les mêmes garanties de diversité quant aux types de texte inclus, ce qui biaiserait les études ultérieures.

Une autre limite est celle de l'étanchéité des communautés concernées. Institutionnellement, en France, le TALN et la linguistique[292] relèvent de deux secteurs disciplinaires aux fonctionnements éloignés : entre ces domaines, les passerelles et les collaborations sont encore fragiles. Les formations autour du traitement automatique du langage, par exemple, relèvent dans l'immédiat d'un secteur ou de l'autre, mais pas d'une convergence des deux.

L'évolution actuelle peut enfin conduire à marginaliser des travaux perçus comme moins directement « utiles ». L'étude diachronique de la langue en fournit un exemple. Mais l'expérimentation de formalismes sophistiqués peut également pâtir du nouveau contexte.

54.3 Questionnements

Du côté linguistique, les travaux que nous avons présentés poussent à examiner, ou à réexaminer sur des bases renouvelées, des phénomènes jusqu'à présent insuffisamment étudiés : place de la ponctuation, structuration globale des textes et grammaires textuelles, articulation langue générale / langues de spécialité, etc.

Du côté informatique, le succès pratique du métissage des traitements à règles et des traitement numériques pose sur le fond la question de modèles qui articulent finement observation et appel à la compétence des locuteurs et à l'expertise des spécialistes.

Une question reste ouverte : quelles généralisations permettent les multiples constats, si fins soient-ils, opérés sur les corpus annotés ?

55. PERSPECTIVES

Sans nous risquer à prédire l'avenir des linguistiques de corpus, nous soulignons à la fois les menaces qui pèsent sur leur développement et les espoirs qui semblent permis. Nous terminons par ce qui nous paraît être les conditions d'une évolution positive du domaine.

[292] Il faudrait en outre mentionner le secteur de l'informatique documentaire, dont les recherches sont mal connues en linguistique et en TALN, bien qu'elles soient riches d'enseignement pour le traitement des corpus annotés.

55.1 Menaces

Les menaces sont de trois ordres : les retards méthodologiques et techniques dans les moyens d'utiliser des corpus annotés, les dimensions laissées dans l'ombre par les linguistiques de corpus, et enfin des impasses intellectuelles.

Les moyens matériels de calcul ne cessent de progresser. Le versant logiciel des traitements de corpus accuse un retard d'autant plus sensible, ce qui retarde d'autant les expérimentations et partant, les avancées théoriques. On sait mémoriser des corpus et des ressources langagières de plus en plus vastes. Malgré des initiatives de mise en convergence, il n'existe pas encore de chaînes de traitement standard pour ces données. La normalisation commence à devenir effective pour les corpus. Elle ne l'est pas encore pour les programmes correspondants, qui restent la plupart du temps expérimentaux. On est encore assez loin de « stations de travail textuelles » qui permettraient d'articuler des traitements diversifiés sur des corpus : étiquetage, correction interactive, parsage, annotation sémantique, décomptes et modélisation ...

Certaines dimensions restent peu abordées en linguistique de corpus. C'est le cas de la textualité en tant que telle[293]. Même les études de Biber, lorsqu'elles caractérisent les types de texte comme des constellations de traits linguistiques, ne rendent pas compte de l'organisation des textes au delà de la phrase, de l'enchaînement des énoncés. La dimension pragmatique s'efface également, en raison de la primauté accordée à la morpho-syntaxe.

Nous avons déjà cité l'adage de G. Sampson (1994, p. 180) : « la linguistique de corpus prend le langage tel qu'il est. » Le piège serait ... de le laisser tel qu'il est, c'est-à-dire de n'introduire aucun déplacement théorique. La manipulation des corpus annotés est lourde. Le déferlement des données peut aussi dérouter, par son intrication complexe de phénomènes multiples[294]. Tout le langage s'engouffre. Le risque est alors un empirisme linguistique radical[295], à fleur de données et sans recul.

Ceux qui mettent au point traitements et outils peuvent être de leur côté tentés par une certaine commisération pour les études proprement linguistiques. Ces dernières ne se confronteraient jamais au langage « réel ».

---

[293] J.-P. Sueur (1982, p. 144) dégage tout de même des pistes et montre des premiers résultats.

[294] C. Filmore et B. Atkins (1994) montrent la complexité de l'analyse du verbe *risk* lorqu'on part, comme eux, de corpus : 1 743 contextes fournis par l'APHB (American Publishing House for the Blind) et de 470 extraits du corpus à la base du dictionnaire COBUILD. Ils comparent les tendances observées dans ces contextes avec le traitement opéré dans dix dictionnaires. Ils insistent sur les choix théoriques comme seuls moyens de s'orienter dans le flux des attestations.

[295] L'expression est de M.-P. Péry-Woodley (1995, p. 216).

55.2 Espoirs

Les recherches dont nous venons de dégager les grands traits renouvellent la dimension empirique et expérimentale de la linguistique, en particulier en ce qui concerne la quantification des faits langagiers.

Pour reprendre les termes de C. Jacquemin, une linguistique véritablement expérimentale est possible. Puisque les corpus et les outils entrent de plus en plus dans le domaine public, les résultats présentés par les recherches sont vérifiables sur les mêmes données ou au contraire amendables par confrontation avec d'autres données. Les faits deviennent un peu plus têtus. Expérimenter, c'est aussi pouvoir construire des modèles, symboliques ou quantitatifs, et les tester sur des données.

Comme l'écrit J. Sinclair (1991, p. 100) : « La langue a l'air assez différente quand on en examine un grand morceau d'un coup. » Les distinctions tranchées s'estompent. Aux différents niveaux de l'analyse linguistique, on peut séparer usuel, exceptionnel et tout à fait improbable. On peut désormais quantifier de nouveaux phénomènes. On peut aussi examiner les corrélations entre des traits linguistiques multiples. Mais il reste à acquérir pour la syntaxe et la sémantique une expérience similaire à celle qui a été développée en analyse statistique du lexique. Elle permettra d'attribuer leur véritable dimension aux résultats obtenus actuellement.

55.3 Conditions

Les linguistiques de corpus se révèleront fructueuses comme domaine de recherche si l'on accepte l'imparfait, c'est-à-dire des ressources toujours « impures », et si s'affirment des collaborations soutenues entre linguistes et informaticiens.

Les corpus annotés comme les outils d'annotation reposent sur des approximations. L'ampleur des moyens à réunir force à des solutions qui, sans être jamais vraiment consensuelles, reposent sur des compromis entre des communautés distinctes et des impératifs techniques multiples. Ces solutions dépendent également de l'usage prévu en aval pour les ressources annotées. Cette imperfection ne constitue pas pour autant un obstacle majeur. Nous l'avons vu, il est souvent possible de « faire des détours » pour isoler les phénomènes visés. Sans doute faut-il aussi abandonner l'horizon, illusoire, de corpus « parfaitement » annotés et d'outils ne faisant pas d'erreur. Pourquoi attendre de la « machine » une cohérence et une perfection que l'annotation manuelle n'atteint pas ?

La collaboration de l'Université de Lancaster et du centre de recherche d'IBM Watson (Black *et al.*, 1993) est exemplaire d'une coopération fructueuse entre les deux communautés concernées au premier chef, la linguistique et le TALN. Les linguistes ont vu leur attention attirée sur des

phénomènes souvent conçus comme marginaux et sur la nécessité de les intégrer dans leur description. Les informaticiens ont appris à modéliser des comportements langagiers plus fins que ceux qu'ils traitaient initialement. Les deux communautés ont l'intérêt le plus vif à coopérer. La constitution de vastes corpus finement annotés et la mise au point des outils nécessaires supposent des recherches informatiques importantes et coûteuses. Les linguistes en bénéfieront. Inversement, seuls des travaux poussés en linguistique descriptive permettent de mieux maîtriser les causalités à l'œuvre : influence des types de textes, jeu entre sous-langages et langue générale, poids du temps, etc. Les informaticiens y trouveront matière à améliorer leurs modèles et leurs techniques. Parce que les corpus lui semblent le moyen de constituer les ressources linguistiques nécessaires à des traitements effectifs, le TALN se confronte désormais à toute la complexité du langage. Disposer de corpus annotés renouvelle les méthodes et les objectifs de la linguistique descriptive. Le foisonnement des recherches témoigne de la vigueur du champ. Il y a probablement une chance historique à saisir : celle d'une coopération enfin fructueuse.

TABLE DES MATIERES

BIBLIOGRAPHIE

ABREVIATIONS UTILISEES

55.3.1.1 Actes

ACL : Association for Computational Linguistics
ANLP : Applied Natural Language Processing
COLING : International Conference on Computational Linguistics
EACL : European Chapter of the Association for Computational Linguistics
EURALEX : International Congress on Lexicography
FRANCIL : Journées du réseau FRANÇais des Industries de la Langue
IJCAI : International Joint Conference in Artificial Intelligence
JADT : Journées de l'Analyse des Données Textuelles
SIGIR : Special Interest Group in Information Retrieval (ACM)

55.3.1.2 Revue

TAL : Traitement Automatique des Langues

55.3.1.3 Association

ACM : Association for Computing Machinery

AARTS J. — Corpus linguistics : an appraisal, in : *Computers in Literary and Linguistic research*, Hamesse J., Zampolli A., Champion-Slatkine, Paris-Genève, 1990, 13–28.

ABEILLE A. — *Les nouvelles syntaxes : grammaires d'unification et analyse du français*, Armand Colin, Paris, 1993.

AGIRRE E., RIGAU G. — Word sense disambiguation using conceptual density, in : *COLING'96*, Copenhague, Danemark, 1996, tm. 1, 16–22.

ALTENBERG B. — Recurrent verb-complement constructions in the London-Lund corpus, in : *English language corpora : design, analysis and exploitation*, Aarts J., de Haan P., Oostdijk N., Rodopi, Amsterdam, 1993, 227–246.

AMSTRONG S. (ed.) — *Using Large Corpora*, The MIT Press, Cambridge, Massachusetts, 1994.

ASSADI H., BOURIGAULT D. — Classification d'adjectifs extraits d'un corpus pour l'aide à la modélisation de connaissances, in : *JADT'95*, 1995.

ATWELL E., HUGHES J., SOUTER C. — Amalgam : Automatic mapping among lexico-grammatical annotation models, in : *The Balancing Act : Combining Symbolic and Statistical Approaches to Language*, Las Cruces, USA, 1994, 11–21.

AUTHIER-REVUZ J. — Méta-énonciation et (dé)figement, in : *La locution en discours*, Martins-Baltar M., ENS de Fontenay/St Cloud, Paris, 1995, 17–40.

BARKEMA H. — Determining the syntactic flexibility of idioms, in : *Creating and using English language corpora*, Fries U., Tottie G., Schneider P., Rodopi, Amsterdam, 1994, 39–52.

BARKEMA H. — Idiomaticy in english NPs, in : *English language corpora : design, analysis and exploitation*, Aarts J., de Haan P., Oostdijk N., Rodopi, Amsterdam, 1993, 257–278.

BARNBROOK G. — *Language and Computers - A practical Introduction to the Computer Analysis of Language*, Edinburgh University Press, Edinburgh, 1996.

BASILI R., DELLA ROCCA M., PAZIENZA M. T. — Contextual word sense tuning and disambiguation, *Applied Artificial Intelligence*, *11*, 1997, 235–262.

BASILI R., PAZIENZA M., VELARDI P. — A « not-so-shallow » parser for collocational analysis, *COLING'94*, 1994, 447–453.

BASILI R., PAZIENZA M., VELARDI P. — Acquisition of selectional patterns in sublanguages, *Machine Translation*, *8*, 1993, 175–201.

BASILI R., PAZIENZA M., VELARDI P. — Semi-automatic extraction of linguistic information for syntactic disambiguation, *Applied Artificial Intelligence*, *7*, 1993, 339–364.

BASILI R., PAZIENZA M., VELARDI P. — What can be learned from raw texts ?, *Machine Translation*, *8*, 1993, 147–173.

BECUE M., BOLASCO S. — Les quasi-segments pour une classification automatique des réponses ouvertes, in : *JADT*, Montpellier, 1993, 310–325.

BENSCH P. A., SAVITCH W. J. — An occurrence-based model of word categorization, *Annals of Mathematics and Artificial Intelligence*, *14*, 1995, 1–16.

BENZECRI J.-P. — *L'analyse des correspondances*, Dunod, 1973.

BENZECRI J.-P. — *La taxinomie*, Dunod, 1973.

BERGOUNIOUX A., LAUNAY M.-F., MOURIAUX R., SUEUR J.-P., TOURNIER M. — *La parole syndicale*, Presses Universitaires de France, Paris, 1982.

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

BIBER D. — Representativeness in corpus design, *Linguistica Computazionale*, *IX-X*, 1994, 377–408.

BIBER D. — *Variation accross speech and writing*, Cambridge University Press, Cambridge, 1988.

BIBER D., FINEGAN E. — Intra-textual variation within medical research articles, in : *Corpus-based research into language*, Ooostdijk N., de Haan P., Rodopi, Amsterdam, 1994, 201–222.

BIBER D., FINEGAN E., ATKINSON D. — ARCHER and its challenges : compiling and exploring a representative corpus of historical english registers, in : *Creating and using English language corpora*, Fries U., Tottie G., Schneider P., Rodopi, Amsterdam, 1994, 1–14.

BLACK E., GARSIDE R., LEECH G., EYES E., MCENERY A., LAFFERTY J., MAGERMAN D., ROUKOS S. — *Statistically-driven computer grammars of English : the IBM/Lancaster approach*, Rodopi, Amsterdam, 1993.

BLACKWELL S. — From dirty data to clean language, in : *English language corpora : design, analysis and exploitation*, Aarts J., de Haan P., Oostdijk N., Rodopi, Amsterdam, 1993, 97–106.

BLANCHE-BENVENISTE C. — *Approches de la langue parlée en français*, Ophrys, Paris, 1997.

BLANK I. — Sentence alignment : methods and implementations, *TAL*, *36*, 1-2, 1995, 81–100.

BOGURAEV B., PUSTEJOVSKY J. (eds.) — Corpus processing for lexical acquisition, The MIT Press, Cambridge, 1996.

BOLASCO S. — Sur différentes stratégies dans une analyse des formes textuelles : une expérimentation à partir de données d'enquête, in : *JADT*, Barcelone, 1992, 69–88.

BOUAUD J., HABERT B., NAZARENKO A., ZWEIGENBAUM P. — Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation à deux modélisations conceptuelles, in : Actes *Ingénierie des connaissances*, Roscoff, 1997, 207–223.

BOURIGAULT D. — Analyse syntaxique locale pour le repérage de termes complexes dans un texte, *TAL*, *34*, 2, 1993.

BRILL E. — Transformation-based error-driven learning and natural language processing : A case study in part-of-speech tagging, *Computational Linguistics*, *21*, 4, 1995, 543–565.

BRISCOE T. — Prospects for practical parsing of unrestricted text : robust statistical parsing techniques, in : *Corpus-based research into language*, Ooostdijk N., de Haan P., Rodopi, Amsterdam, 1994, 97–120.

BRONCKART J.-P., BAIN D., SCHNEUWLY B., DAVAUD C., PASQUIER A. — *Le fonctionnement des discours : un modèle psychologique et une méthode d'analyse*, Delachaux & Niestlé, Lausanne, 1985.

BROWN P., LAI J., MERCER R. — Aligning sentences in parallel corpora, in : *ACL'91*, Berkeley, USA, 1991.

BRUNET E. — *Le Vocabulaire de Marcel Proust*, Slatkine-Champion, Genève-Paris, 1983.

BRUNET E. — *Le vocabulaire français de 1789 à nos jours, d'après les données du* Trésor de la langue française, Slatkine-Champion, Genève-Paris, 1981.

BRUNET E. — What do statitistics tell us, in : *Research in humanities Computing*, Clarendon Press, Oxford, tm. 1, 1991, 35–46.

BURNAGE G., DUNLOP D. — Encoding the British National Corpus, in : *English language corpora : design, analysis and exploitation*, Aarts J., de Haan P., Oostdijk N., Rodopi, Amsterdam, 1993, 79–96.

BURNARD L. — *Users Reference Guide for the British National Corpus*, British National Corpus Consortium, Oxford University Computing Services, Oxford, UK, may 1995.

BURNARD L. — What is SGML and how does it help ?, *Computers and the Humanities*, 29, 1995, 41–50.

BURNARD L., SPERBERG-MCQUEEN C. M. — La TEI simplifiée : une introduction au codage des textes électroniques en vue de leur échange, *Cahiers Gutenberg*, 24, 1996, 23–151.

CALLIOPE (COLLECTIF). — *La parole et son traitement automatique*, Masson, Paris, 1989.

CHANOD J.-P., TAPANAINEN P. — Creating a tagset, lexicon and guesser for a french tagger, in : *Proceedings of EACL SIGDAT workshop on From Texts To Tags: Issues In Multilingual Language Analysis*, 1995, 58–64.

CHANOD J.-P., TAPANAINEN P. — Tagging French — comparing a statistical and a constraint-based method, in : *EACL'95*, Dublin, 1995, 149–156.

CHARLET J., BACHIMONT B., BOUAUD J., ZWEIGENBAUM P. — Ontologie et réutilisabilité : expérience et discussion, in : *Acquisition et ingénierie des connaissances : tendances actuelles*, Aussenac-Gilles N., Laublet P., Reynaud C., Cépaduès Editions, Toulouse, 1996, 69–87.

CHISHOLM D., ROBEY D. — Encoding verse texts, *Computers and the Humanities*, 29, 1995, 99–111.

CHURCH K. W. — Char Align: A program for aligning parallel texts at the character level, in : *ACL'93*, Columbus, Ohio, 1993.

CHURCH K. W. — One term or two ?, in : *SIGIR*, Seattle, USA, 1995, 310–318.

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

CHURCH K. W., MERCER R. L. — Introduction to the special issue on Computational Linguistics Using Large Corpora, *Computational Linguistics*, *19*, 1, 1993, 1–24.

CHURCH K., GALE W. — Concordance for Parallel Texts, in : *Proceedings of the 7th Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, Oxford, 1991.

COVER R. C., ROBINSON P. M. W. — Encoding textual criticism, *Computers and the Humanities*, 29, 1995, 123–136.

COWIE J., GUTHRIE J., GUTHRIE L. — Lexical disambiguation using simulated annealing, in : *COLING'92*, Nantes, 1992, 359–365.

CUTTING D., KUPIEC J., PEDERSEN J., SIBUN P. — A practical part-of-speech tagger, in : *ANLP'92*, 1992.

DAGAN I., ITAI A., SCHWALL U. — Two languages are more informative than one, in : *ACL'91*, Berkeley, USA, 1991, 130–137.

DAILLE B. — Repérage et extraction de terminologie par une approche mixte statistique et linguistique, *TAL*, *36*, 1-2, 1995, 101–118.

DAILLE B. — Study and implementation of combined techniques for automatic extraction of terminology,in : Actes *The Balancing Act - Combining Symbolic and Statistical Approaches to Language*, Las Cruces, USA, 1995, 29–36.

DALADIER A. — Aspects constructifs des grammaires de Harris, *Langages*, 99, 1990, 57–84.

DUNLOP D. — Practical considerations in the use of TEI headers in large corpora, *Computers and the Humanities*, 29, 1995, 85–98.

DUPUIS F., LEMIEUX M., GOSSELIN D. — Conséquences de la sous-spécification des traits de Agr dans l'identification de Pro, *Language Variation and Change*, 3, 1992, 275–299.

EEG-OLOFSSON M., ALTENBERG B. — Discontinuous recurrent word combinations in the London-Lund corpus, in : *Creating and using English language corpora*, Fries U., Tottie G., Schneider P., Rodopi, Amsterdam, 1994, 63–78.

EL-BÈZE M., SPRIET T. — Intégration de contraintes syntaxiques dans un système d'étiquetage probabiliste, *TAL*, *36*, 1-2, 1995, 47–66.

ENGWALL G. — Not chance but choice : Criteria in corpus creation, in : *Computational Approaches to the Lexicon*, Atkins B., Zampolli A., Oxford University Press, Oxford, 1994, 49–82.

EVANS D. A., ZHAI C. — Noun-phrase analysis in unrestricted text for information retrieval, in : *ACL'96*, Santa Cruz, USA, 1996.

EYES E., LEECH G. — Progress in UCREL research : improving corpus annotation practices, in : *English language corpora : design, analysis and exploitation*, Aarts J., de Haan P., Oostdijk N., Rodopi, Amsterdam, 1993, 125–143.

FELLBAUM C., GROSS D., MILLER K. — Adjectives in WordNet, in : *Five Papers on WordNet*, http://www.cogsci.princeton.edu/ wn/ (sept. 1997), 1993, 26–39, revised version.

FIALA P., HABERT B. — La langue de bois en éclats : les défigements dans les titres de la presse quotidienne française, *MOTS,* 1989, 83–98.

FILLMORE C. J., ATKINS B. — Starting where the dictionaries stop : The challenge of corpus lexicography, in : *Computational Approaches to the Lexicon*, Atkins B., Zampolli A., Oxford University Press, Oxford, 1994, 349–396.

FINEGAN E., BIBER D. — *That* and zero complementisers in late modern english : exploring archer from 1650-1990, in : *The verb in contemporary English. Theory and description*, Aarts B., Meyer C. F., Cambridge University Press, Cambridge, 1995, 241–257.

FUCHS C. (resp.) — *Linguistique et traitement automatique des langues*, Hachette, Paris, 1993.

GALE W. A., CHURCH K. W. — A program for aligning sentences in bilingual corpora, *Computational Linguistics*, *19*, 1, 1993, 75–102.

GAUSSIER E., GREFENSTETTE G., SCHULZE M. — Traitement du langage naturel et recherche d'information : quelques expériences sur le français, in : *FRANCIL'97*, 1997, 9–14.

GAUSSIER E., LANGE J.-M. — Modèles statistiques pour l'extraction de lexiques bilingues, *TAL*, *36*, 1-2, 1995, 133–156.

GAZDAR G., KLEIN E., PULLUM G. K., SAG I. A. — *Generalized Phrase Structure Grammar*, Harvard University Press, Cambridge, MA, 1985.

GAZDAR G., MELLISH C. — *Natural Language Processing in Lisp*, Addison Wesley, Reading, 1989.

GIORDANO R. — The TEI header and the documentation of electronic texts, *Computers and the Humanities*, 29, 1995, 75–85.

GOLDFARB C. F. — *The SGML Handbook*, Clarendon Press, 1990.

GOOSSENS M. — Introduction pratique à SGML, *Cahiers Gutenberg*, 19, 1995, 27–58.

GRANGER S. — International corpus of learner english, in : *English language corpora : design, analysis and exploitation*, Aarts J., de Haan P., Oostdijk N., Rodopi, Amsterdam, 1993, 57–71.

GREENBAUM S. — The tagset for the International Corpus of English, in : *Corpus-Based Computational Linguistics*, Souter C., Atwell E., Rodopi, Amsterdam, 1993, 11–24.

GREENBAUM S., YIBIN N. — Tagging the British ICE corpus : English word classes, in : *Corpus-based research into language*, Ooostdijk N., de Haan P., Rodopi, Amsterdam, 1994, 33–46.

GREENSTEIN D., BURNARD L. — Speaking with one voice : Encoding standards and the prospects for an integrated approach to computing in history, *Computers and the Humanities*, 29, 1995, 137–148.

GREFENSTETTE G. — Automatic thesaurus generation from raw text using knowledge-poor techniques, in : *Proceedings of the 9th Conference on Oxford English dictionary*, Oxford, 1993.

GREFENSTETTE G. — Corpus-derived first, second and third order affinities, in : *EURALEX*, Amsterdam, 1994.

GREFENSTETTE G. — Evaluation techniques for automatic semantic extraction : Comparing syntactic and window based approaches, in : *Corpus Processing for Lexical Acquisition*, Boguraev B., Pustejovsky J., The MIT Press, Cambridge, Massachusetts, 1996, 205–216.

GRISHMAN R., KITTREDGE R., (eds.): *Analyzing Language in Restricted Domains. Sublanguage Description and Processing.*, Lawrence Erlbaum Ass., Hillsdale, 1986.

GRISHMAN R., STERLING J. — Generalizing automatically generated selectional patterns, in : *COLING'94*, Kyoto, 1992, tm. 3, 742–747.

GROSS G. — Classes d'objets et description des verbes, *Langages*, 115, 1994, 15–30.

GROSS G. — Degré de figement des noms composés, *Langages*, 90, 1988, 57–70.

GUHA R., LENAT D. B. — Enabling agents to work together, *Communications of the ACM*, *37*, 7, 1994, 127–142.

GUILLET A. — Fondements formels des classes sémantiques dans un lexique-grammaire, *Langages 98*, 1990, 70–102.

GUTHRIE J., GUTHRIE L., WILKS Y., AIDINEJAD H. — Subject-dependent co-occurrences and word sense disambiguation, in : *ACL'91*, Berkeley, USA, 1991.

HABERT B. (resp.) — Traitements probabilistes et corpus, *TAL,* 36, 1-2, 1995.

HABERT B. — Études des formes spécifiques et typologie des énoncés (les résolutions générales des congrès de la CFTC-CFDT de 1945 à 1979), *MOTS,* 11, 1985, 127–154.

HABERT B. — L'analyse des formes spécifiques. Bilan critique et propositions d'utilisation, *MOTS,* 7, 1983, 97–124.

HABERT B., HERVIOU-PICARD M.-L., BOURIGAULT D., QUATRAIN R., ROUMENS M. — Un outil et une méthode pour comparer deux extracteurs de groupes nominaux, in : *FRANCIL'97*, 1997, 509-516.

HABERT B., NAULLEAU E., NAZARENKO A. — Symbolic word clustering for medium-size corpora, in : *COLING'96*, Copenhague, Danemark, 1996, tm. 1, 490–495.

HABERT B., SALEM A. — L'utilisation de catégorisations multiples pour l'analyse quantitative de données textuelles, *TAL*, *36*, 1-2, 1995, 249–276.

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, 1989.

HATZIVASSILOGLOU V., MCKEOWN K. — Towards the automatic identification of scales : Clustering adjecives according to meaning, in : *ACL'93*, Columbus, USA, june 1993, 172–182.

HEARST M. A. — Automatic acquisition of hyponyms from large text corpora, in : *COLING'92*, Nantes, 1992, 539–545.

HERDAN G. — *Quantitative Linguistics*, Butterworths, Londres, 1964.

HERZOG O., ROLLINGER C. (eds.): *Text Understanding in LILOG*, Springer-Verlag, Heidelberg, 1991.

HINDLE D. — A parser for text corpora, in : *Computational Approaches to the Lexicon*, Atkins B., Zampolli A., Oxford University Press, Oxford, 1994, 103–152.

HINDLE D. — Noun classification from predicate argument structures, in : *ACL'83*, Berkeley, USA, 1990, 268–275.

HOLMES D. I. — The analysis of literary style - A review, *J.R. Statistic. Soc.*, 148, Part 4, 1985, 328–341.

HUMPHREY B. L., LINDBERG D. A. — Building the Unified Medical Language System, in : *Proceedings of the 6th Annual SCAMC*, IEEE, Washington, 1989, 475–480.

IDE N., SPERBERG-MCQUEEN C. M. — The Text Encoding Initiative — its history, goals and future development, *Computers and the Humanities*, 29, 1995, 5–16.

IDE N., VÉRONIS J. (eds.) — *The Text Encoding Initiative: Background and context*, Kluwer Academic Publishers, Dordrecht, 1995.

IDE N., VÉRONIS J. — Encoding dictionaries, *Computers and the Humanities*, 29, 1995, 167–180.

ISABELLE P. — La bi-textualité : vers une nouvelle génération d'aides à la traduction et à la terminologie, *META*, *37*, 4, 1992, 721–737.

ISABELLE P., WARWICK-ARMSTRONG S. — Les corpus bilingues : une nouvelle ressource pour le traducteur, in : *La traductique*, Bouillon P., Clas A., Presses de l'Université de Montréal, Montréal, 1993, 288–306.

JACQUEMIN C., KLAVANS J. L., TZOUKERMANN E. — Expansion of multi-word terms for indexing and retrieval using morphology and syntax, in : *ACL - EACL'97*, Madrid, 1997, 24–31.

JACQUEMIN C., ROYAUTÉ J. — Retrieving terms and their variants in a lexicalized unification-based framework, in : *SIGIR'94*, Dublin, 1994, 132–141.

JOHANSSON S. — 'This scheme is badly needed' : some aspects of verb-adverb combinations, in : *The verb in contemporary English. Theory and description*, Aarts B., Meyer C. F., Cambridge University Press, Cambridge, 1995, 218–240.

JOHANSSON S. — Continuity and change in the encoding of computer corpora, in : *Corpus-based research into language*, Ooostdijk N., de Haan P., Rodopi, Amsterdam, 1994, 13–32.

JOHANSSON S. — The encoding of spoken texts, *Computers and the Humanities*, 29, 1995, 149–158.

JUSTESON J. S., KATZ S. M. — Principled disambiguation : Discriminating adjective senses with modified nouns, *Computational Linguistics*, *21*, 1, 1995, 1–28.

KARLSSON F. — Robust parsing of unconstrained text, in : *Corpus-based research into language*, Oostdijk N., de Haan P., Rodopi, Amsterdam, 1994, 121–142.

KARLSSON F., VOUTILAINEN A., HEIKKILA J., ANTILLA A. — *Contraint Grammar : a Language-Independent System for Parsing Unrestricted Text*, Mouton de Gruyter, 1995.

KLEIBER G. — Dénomination et relations dénominatives, *Langages*, 76, 1984, 77–94.

KROCH A. S. — Reflexes of grammar in patterns of language change, *Language Variation and Change*, 3, 1990, 275–299.

KROVETZ R. — Lexical acquisition and information retrieval, in : *Lexical Acquisition : Exploiting On-Line Ressources to build a Lexicon*, Zernik U., Lawrence Erlbaum, USA, 1991.

KUCERA H., NELSON F. — *Computational Analysis of Present-Day American English*, Brown University Press, Providence, 1967.

KYTÖ M. — A supplement to the Helsinki corpus of english texts : the corpus of early american english, in : *English language corpora : design, analysis and exploitation*, Aarts J., de Haan P., Oostdijk N., Rodopi, Amsterdam, 1993, 289–298.

LABBE D. — *Le vocabulaire de François Mitterrand*, Presses de la Fondation Nationale des Sciences Politiques, Paris, 1990.

LAFON P. — Analyse lexicométrique et recherche des cooccurrences, *MOTS*, 3, 1981, 95–148.

LAFON P. — Sur la variabilité de la fréquence des formes dans un corpus, *MOTS*, 1, 1980, 128–165.

LAFON P., SALEM A. — L'inventaire des segments répétés d'un texte, *Mots*, 6, 1983, 161–177.

LANGE J.-M., GAUSSIER E. — Alignement de corpus multilingues au niveau des phrases, *TAL*, *36*, 1-2, 1995, 67–80.

LAVAGNINO J., MYLONAS E. — The show must go on : Problems of tagging performance texts, *Computers and the Humanities*, 29, 1995, 113–121.

LE PESANT D. — Les compléments nominaux du verbe *lire* : une illustration de la notion de « classe d'objets », *Langages*, 115, septembre 1994, 31–46.

LEBART L., SALEM A. — *Statistique textuelle*, Dunod, Paris, 1994.

LEECH G. — The state of the art in corpus linguistics, in : *English Corpus Linguistics*, Aijmer K., Altenberg B., Longman, London, 1991, 8–29.

LEECH G., BARNETT R., KAHREL P. — *Preliminary recommendations for the Syntactic Annotation of Corpora*, Rap. tech., EAGLES (Expert Advisory Group on Language Engineering Standards), march 1996, CEE.

LEECH G., BARNETT R., KAHREL P. — *Syntactic Annotation : Survey of Annotation Practices*, Rap. tech., EAGLES (Expert Advisory Group on Language Engineering Standards), april 1995, CEE.

LEECH G., GARSIDE R., ATWELL E. — The automatic grammatical tagging of the LOB corpus, *Newsletter of the International Computer Archive of Modern English*, *7*, 1983, 13–33.

LEECH G., GARSIDE R., BRYANT M. — The large-scale grammatical tagging of text : experience with the British National Corpus, in : *Corpus-based research into language*, Ooostdijk N., de Haan P., Rodopi, Amsterdam, 1994, 47–64.

LIBERMAN M. Y. — The Trend towards Statistical Models in Natural Language Processing, in : *Natural Language and Speech*, Klein E., F. Veltman, Springer-Verlag, 1991, 1-7.

LIGOZAT G. — *Représentation des connaissances et linguistique*, Armand Colin, Paris, 1994.

MAINGUENEAU D. — *L'analyse du discours : introduction aux lectures de l'archive*, Hachette, Paris, 1991.

MAIR C. — Changing patterns of complementation, and concomitant grammaticalisation, of the verb *help* in present-day british english, in : *The verb in contemporary English. Theory and description*, Aarts B., Meyer C. F., Cambridge University Press, Cambridge, 1995, 258–271.

MAIR C. — Is *see* becoming a conjunction ? the study of grammaticalisation as a meeting ground for corpus linguistics and grammatical theory, in : *Creating and using English language corpora*, Fries U., Tottie G., Schneider P., Rodopi, Amsterdam, 1994, 127–137.

MARANDIN J.-M., CORI M. — Grammaires d'arbres polychromes, *TAL.*, *34*, 1, 1993, 101–132.

MARANDIN J.-M. — Analyseurs syntaxiques. Equivoques et problèmes, *TAL, 34*, 1, 1993, 5–34.

MARCHELLO-NIZIA C. — *L'évolution du français : ordre des mots, démonstratifs, accent tonique*, Armand Colin, Paris, 1995.

MARCUS M. P., HINDLE D., FLECK M. M. — D-theory : Talking about talking about trees, in : *ACL'83*, 1983, 129–136.

MARCUS M., SANTORINI B., MARCINKIEWICZ M. A. — Building a large annotated corpus of english : The Penn Treebank, *Computational Linguistics*, *19*, 2, 1993, 313–330.

MATHIEU-COLAS M. — *Les mots à traits d'union. Problèmes de lexicographie informatique*, Paris, 1994.

MCENERY T., WILSON A. — *Corpus Linguistics*, Edinburgh University Press, Edinburgh, 1996.

MCMAHON J. G., SMITH F. J. — Improving statistical language model performance with automatically generated word hierarchies, *Computational Linguistics*, *22*, 2, 1996, 217–247.

MCNaught J. — User needs for textual corpora in natural language processing, *Literary and Linguistic Computing*, *8*, 9, 1993, 227–234.

MEL'CUK I. — Paraphrase et lexique dans la théorie linguistique sens-texte, *Lexique*, 6, 1988, 13–54.

MELBY A. — E-TIF : an electronic terminology interchange format, *Computers and the Humanities*, 29, 1995, 159–166.

MELIS-PUCHULU A. — Les adjectifs dénominaux : des adjectifs de « relation », *Lexique*, 10, 1991, 33–60.

MERIALDO B. — Modèles probabilistes et étiquetage automatique, *TAL*, *36*, 1-2, 1995, 7–22.

MERIALDO B. — Tagging english text with a probabilistic model, *Computational Linguistics*, *20*, 2, 1994, 155–171.

MILKHEEV A., FINCH S. P. — A workbench for acquisition of ontological knowledge from natural language, in : *Actes, 9th Knowledge Acquisition for Knowledge-Based Systems Workshop*, Banff, 1995.

MILLER G. A. — Nouns in WordNet : A lexical inheritance system, in : *Five Papers on WordNet*, http://www.cogsci.princeton.edu/ wn/ (sept. 1997), 1993, 10–25, revised version.

MILLER G. A., BECKWITH R., FELLBAUM C., GROSS D., MILLER K. J. — Introduction to WordNet¸: An on-line lexical database, *Journal of Lexicography*, *3*, 1990, 235–244.

MILLER G. A., BECKWITH R., FELLBAUM C., GROSS D., MILLER K. — Introduction to WordNet : An on-line lexical database, in : *Five Papers on WordNet*, http://www.cogsci.princeton.edu/ wn/ (sept. 1997), 1993, 1–9, revised version.

MILNER J.-C. — *Introduction à une science du langage*, Des Travaux, Seuil, Paris, 1e édn., 1989.

MULLER C. — *Initiation aux méthodes de la statistique linguistique*, Hachette, Paris, 1973.

NEDERHOF M. J., KOSTER K. — A customized grammar workbench, in : *English language corpora : design, analysis and exploitation*, Aarts J., de Haan P., Oostdijk N., Rodopi, Amsterdam, 1993, 163–180.

NEVALAINEN T. — Diachronic issues in english adverb derivation, in : *Creating and using English language corpora*, Fries U., Tottie G., Schneider P., Rodopi, Amsterdam, 1994, 139–147.

NUNBERG G. — *The Linguistics of Punctuation*, CSLI, Menlo Park, 1990.

PAROUBEK P., ADDA G., MARIANI J., RAJMAN M. — Les procédures de mesure automatique de l'action GRACE pour l'évaluation des assignateurs de parties du discours pour le français, in : *FRANCIL'97*, Avignon, 1997, 245–252.

PARTEE B. H., MEULEN A. T., WALL R. E. — *Mathematical models in linguistics*, Kluwer Academic Publishers, 1990.

PÊCHEUX M. — *Analyse automatique du discours*, Dunod, Paris, 1969.

PEREIRA F., TISHBY N., LEE L. — Distributional clustering of english words, in : *ACL'93*, Columbus, USA, 22-26 june 1993, 183–190.

PERY-WOODLEY M.-P. — Quels corpus pour quels traitements automatiques ?, *TAL*, *36*, 1-2, 1995, 213–232.

PERY-WOODLEY M.-P. — *Les écrits dans l'apprentissage : clés pour analyser les productions des apprenants*, F References, Hachette, Paris, 1993.

PETERS P. — American and british influence in australian verb morphology, in : *Creating and using English language corpora*, Fries U., Tottie G., Schneider P., Rodopi, Amsterdam, 1994, 149–158.

PUJOL N. — *Corpora : éléments pour un Guide Juridique*, Rap. tech., Institut de Recherches Comparatives sur les Institutions et le Droit - CNRS, Ivry-sur-Seine, 1993.

QUIRK R., GREENBAUM S., LEECH G., SVARTVIK J. — *A Comprehensive Grammar of the English Language*, Longman, London, 1985.

RAJMAN M. — Approche probabiliste de l'analyse syntaxique, *TAL*, *36*, 1-2, 1995, 157–201.

RAUMOLIN-BRUNBERG H. — The position of adjectival modifiers in late middle english noun phrases, in : *Creating and using English language corpora*, Fries U., Tottie G., Schneider P., Rodopi, Amsterdam, 1994, 159–168.

REINERT M. — Alceste, une méthodologie d'analyse des données textuelles et une application : Aurélia de Gérard de Nerval, *Bull. de Méthod. Sociol.*, 26, 1990, 24–54.

RENOUF A. — A word in time : first findings from the investigation of dynamic text, in : *English language corpora : design, analysis and exploitation*, Aarts J., de Haan P., Oostdijk N., Rodopi, Amsterdam, 1993, 279–288.

RESNIK P. — Disambiguation noun groupings with respect to WordNet senses, in : *Third Workshop on Very Large Corpora*, Yarowsky D., Church K., Cambridge, USA, 1995, 54–68.

RESNIK P. — Using information content to evaluate semantic similarity in a taxonomy, in : *IJCAI'95*, 1995.

REY A., CHANTREAU S. — *Dictionnaire des expressions et locutions*, Le Robert, Paris, 1979.

RILOFF E. — Little words can make a big difference for text classification, in : *SIGIR*, Seattle, USA, 1995, 130–136.

ROLE F. — Le codage informatique des apparats critiques : évaluation des recommandations de la Text Encoding Initiative, *Cahiers Gutenberg*, 24, juin 1996, 153–165.

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

SAGER N., FRIEDMAN C. (eds.) — *Medical Language Processing : Computer Management of Narrative Data*, Addison-Wesley, Reading, 1987.

SALEM A. — *Pratique des segments répétés : essai de statistique textuelle*, Kliencksieck, Paris, 1987.

SALTON G. — *Automatic Text Processing¸: The Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley, Reading, 1989.

SAMPSON G. — Susanne : a domesday book of english grammar, in : *Corpus Based Research into Language*, Oostdijk N., de Haan P., Rodopi, Amsterdam, 1994, 169–187.

SAPORTA G. — *Probabilités analyse des données et statistique*, Technip, Paris, 1990.

SCHMIED J. — Analysing style variation in the east african corpus of english, in : *Creating and using English language corpora*, Fries U., Tottie G., Schneider P., Rodopi, Amsterdam, 1994, 167–174.

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

SIMARD M., FOSTER G., ISABELLE P. — Using cognates to align sentences in bilingual corpora, in : *Proc. of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 92)*, Montreal, Canada, 1992.

SIMONIN-GRUMBACH J. — Pour une typologie des discours, in : *Langue, discours, société (pour Emile Benveniste)*, Seuil, Paris, 1975, 85–121.

SINCLAIR J. — *Preliminary recommendations on Corpus Typology*, Rap. tech., EAGLES (Expert Advisory Group on Language Engineering Standards), may 1996, CEE.

SINCLAIR J., HANKS P., FOX G., MOON R., STOCK P. (eds.): *Collins COBUILD English Language Dictionary*, Collins, Glasgow, 1987.

SMADJA F. — Retrieving collocations from text: Xtract, *Computational Linguistics*, *19*, 1, 1993, 143–177.

SOUTER C. — Towards a standard format for parsed corpora, in : *English language corpora : design, analysis and exploitation*, Aarts J., de Haan P., Oostdijk N., Rodopi, Amsterdam, 1993, 197–212.

SOUTER C., ATWELL E. — Using parsed corpora : a review of current practice, in : *Corpus-based research into language*, Ooostdijk N., de Haan P., Rodopi, Amsterdam, no. 12 dans Language and computers : studies in practical linguistics, 1994, 143–158.

SRINIVASAN P. — Thesaurus construction, in : *Information Retrieval : Data Structures and Algorithms*, Frakes W. B., Baeza-Yates R., Prentice Hall, New Jersey, 1992.

STEIN A., SCHMID H. — Étiquetage morphologique de textes français avec un arbre de décision, *TAL*, *36*, 1-2, 1995, 23–36.

SUEUR J.-P. — Pour une grammaire du discours : élaboration d'une méthode; exemples d'application, *MOTS*, 5, 1982, 145–185.

SUSSNA M. — Word sense disambiguation for free-text indexing using a massive semantic network, in : *Proceedings of the Second International Conference on Information and Knowledge Management*, Bhargava B., Finin T., Yesha Y., ACM, 1993, 67–74.

SVARTVIK J., EEG-OLOFSSON M., FORSHEDEN O., ORESTRÖ B., THAVENIUS C. — *Survey of Spoken English*, Lund University Press, Lund, 1982.

TAPANAINEN P., JÄRVINEN T. — Syntactic analysis of natural language using linguistic rules and corpus-based patterns, in : *EACL'95*, Dublin, 1995.

TODOROV T. — *M. Bakhtine, Le principe dialogique*, Le Seuil, Paris, 1981.

TZOUKERMANN E., RADEV D. R. — Using word class for part-of-speech disambiguation, in : *Fourth Workshop on Very Large Corpora*, Ejerhed E., Dagan I., Copenhague, Danemark, 1996, 1–13.

USHIODA A. — Hierarchical clustering of words and application to nlp tasks, in : *4th Workshop on Very Large Corpora*, Ejerhed E., Dagan I., Copenhague, Danemark, 1996, 28–41.

VAN HALTEREN H., DEN HEUVEL T. V. — *Linguistic exploitation of syntactic databases : the use of the Nijmegen Linguistic DataBase program*, Rodopi, Amsterdam, 1990.

VAN HALTEREN H., OOSTDIJK N. — Towards a syntactic database : the TOSCA analysis system, in : *English language corpora : design, analysis and exploitation*, Aarts J., de Haan P., Oostdijk N., Rodopi, Amsterdam, 1993, 145–162.

VAN HERWIJNEN E. — *SGML pratique*, International Thomson Publishing France, Paris, 1995.

VAN DER LINDER E. J. — Incremental processing and the hierarchical lexicon, *Computational Linguistics*, *18*, 2, 1992, 218–237.

VÉRONIS J., IDE N. — Word sense disambiguation with very large neural networks extracted from machine readable dictionaries, in : *COLING'90*, Helsinki, Finlande, 1990, 389–394.

VERONIS J., KHOURI L. — Étiquetage grammatical multilingue : le projet MULTEXT, *TAL*, *36*, 1-2, 1995, 233–248.

VIJAY-SHANKER K. — Using descriptions of trees in a Tree Adjoining Grammar, *Computational Linguistics*, *18*, 4, 1992, 482–516.

VOORHEES E. M. — Query expansion using lexical-semantic relations, in : *SIGIR'94*, 1994.

VOSSEN P. — Right or wrong : Combining lexical resources in the EuroWordNet project, in : *EURALEX '96*, Suède, 1996, tm. II, 715–728.

VOUTILAINEN A., HEIKKILÄ J. — An english constraint grammar (ENGCG): a surface-syntactic parser of english, in : *Creating and using English language corpora*, Fries U., Tottie G., Schneider P., Rodopi, Amsterdam, 1994, 189–200.

WARNESSON I. — Applied linguistics : optimization of semantic relations by data aggregation techniques, *Applied Stochastic Models and Data Analysis*, *1*, 1985, 121–141.

WRIGHT S. — In search of history : English language in the eighteenth century, in : *English language corpora : design, analysis and exploitation*, Aarts J., de Haan P., Oostdijk N., Rodopi, Amsterdam, 1993, 25–39.

WRIGHT S. — The place of genre in corpus, in : *Corpora across the centuries*, Kytö M., Rissanen M., Wright S., Rodopi, Amsterdam, 1994, 101–110.

YAROWSKY D. — Word-sense disambiguation using statistical models of Roget's categories trained on large corpora, in : *COLING'92*, Nantes, 1992, p. 454460.

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

INDEX

Cet ouvrage présente un panorama de travaux récents dans le domaine du traitement automatique des textes. L'ouvrage décrit les principaux types de ressources informatisées actuellement disponibles : corpus de textes ayant fait l'objet d'annotations morphologiques, syntaxiques ou sémantiques, ressources dictionnairiques, procédures permettant d'enrichir automatiquement ou semi-automatiquement des textes réunis en corpus. L'utilisation conjointe de ces ressources est illustrée à partir d'exemples empruntés à des recherches effectives menées dans des domaines très divers. Au-delà de la communauté des linguistes et de celle du traitement automatique du langage, cet ouvrage concerne les lexicographes, les didacticiens, les analystes de contenu, etc., ainsi que tous ceux que leur travail confronte à l'étude de la langue, du discours et des textes.

***Benoît Habert**, ancien élève de l'ENS de Saint-Cloud, agrégé de lettres modernes, docteur de 3e cycle en linguistique et docteur en informatique, est maître de conférences en informatique à l'ENS de Fontenay-Saint-Cloud.*

***Adeline Nazarenko**, ancienne élève de l'ENS, agrégée de lettres modernes, docteur en informatique, est maître de conférences en informatique à l'université Paris-XIII.*

***André Salem**, docteur de 3e cycle en statistique mathématique, docteur d'État ès lettres et sciences humaines, est maître de conférences en sciences du langage à l'université Paris-III.*

Cet ouvrage présente un panorama de travaux récents dans le domaine du traitement automatique des textes. L'ouvrage décrit les principaux types de ressources informatisées actuellement disponibles : corpus de textes ayant fait l'objet d'annotations morphologiques, syntaxiques ou sémantiques, ressources dictionnairiques, procédures permettant d'enrichir automatiquement ou semi-automatiquement des textes réunis en corpus. L'utilisation conjointe de ces ressources est illustrée à partir d'exemples empruntés à des recherches effectives menées dans des domaines très divers. Au-delà de la communauté des linguistes et de celle du traitement automatique du langage, cet ouvrage concerne les lexicographes, les didacticiens, les analystes de contenu, etc., ainsi que tous ceux que leur travail confronte à l'étude de la langue, du discours et des textes.

***Benoît Habert**, ancien élève de l'ENS de Saint-Cloud, agrégé de lettres modernes, docteur de 3e cycle en linguistique et docteur en informatique, est maître de conférences en informatique à l'ENS de Fontenay-Saint-Cloud.*

***Adeline Nazarenko**, ancienne élève de l'ENS, agrégée de lettres modernes, docteur en informatique, est maître de conférences en informatique à l'université Paris-XIII.*

***André Salem**, docteur de 3e cycle en statistique mathématique, docteur d'État ès lettres et sciences humaines, est maître de conférences en sciences du langage à l'université Paris-III.*