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

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

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.

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

[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.

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 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

[5] C'est le cas du *Survey of English Usage* de R. Quirk et de (Quirk *et al.*, 1985).
[6] Intervention radio-télévisée du 2 mars 1986.

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 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].

[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. »

[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

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 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

---

statistiques et probabilistes ... pour dire qu'ils ne seraient pas abordés.

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 repris dans l'index.

Les crochets servent à isoler des suites de traits linguistiques, qui sont mis entre accolades : [{nom commun}{adjectif relationnel}] désigne

[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.

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

[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.

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 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

[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).

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.

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

[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. »

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 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).