CHAPITRE VII

CONSTITUER UN CORPUS

1. 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[1], 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 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

[1] Ce qui est appelé crûment dans (Marcus *et al.*, 1993, p. 313, n. 1) des regroupements « opportunistes ».

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[2]. 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[3] (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[4]. » 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 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

[2] 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.

[3] 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 ...

[4] 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. »

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*[5] – *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).

2. LANGUE GENERALE

2.1 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

[5] ou encore *corpus baromètre*.

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.

2.2 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[6]. 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.

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[7], 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

[6] 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.

[7] 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).

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).

2.3 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 représentativité[8]. 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).

3. 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,

[8] On se reportera à (Biber, 1993a, 1994) pour une discussion approfondie.

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.

3.1 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[9] dans ce qu'il appelle les sous-langages[10] : 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 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[11] ressortit à la classe des médecins et N2 à celle des artères : on dilate une artère coronaire, une artère circonflexe, etc[12].

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 »[13]. 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

[9] « 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).

[10] (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).

[11] N1 n'est pas toujours exprimé, par exemple dans la nominalisation *dilatation de N2* ou dans l'utilisation du passif *N2 a été dilaté*.

[12] Il s'agit d'ailleurs d'une métonymie, c'est en fait un segment qui est dilaté et non l'artère entière.

[13] 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).

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).

3.2 Analyses de sous-langages

3.2.1 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 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[14] 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.

[14] 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.

3.2.2 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.

3.3 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 générale. En outre, l'accent porte sur une caractérisation avant tout syntaxique : la sémantique est conçue comme trop peu formalisable[15], 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

[15] 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.

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 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 ?

4. 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.

4.1 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.

4.2 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.

5. 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.

5.1 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[16]. C'est l'équivalent de la transformation que nous avons opérée lors de la présentation de l'étiquetage lorsque nous avons 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+[17]

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>
```

[16] N. Ide et J. Véronis (1995b) analysent en détail le codage des dictionnaires.

[17] 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.

```
    <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>
  </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[18] qui est maintenant présent dans pratiquement tout logiciel de gestion de document[19]. 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[20].

5.2 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

[18] 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 ...).

[19] 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.

[20] (van Herwijnen, 1994) constitue une introduction globale et pratique à SGML.

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[21] *(Text 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[22] en 1994. De nombreux projets de constitution de corpus et de ressources linguistiques ont adopté la TEI (***BNC*** par exemple)[23]. 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.

6. 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[24].

---

[21] 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 la CEE, la fondation Andrew W. Mellon et le Social Science and Humanities Research Council du Canada.

[22] La TEI est donc une proposition de norme et non une norme.

[23] 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.

[24] ***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).

6.1 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.

6.2 Jurisprudence d'annotation

La qualité primordiale d'un système d'annotation, c'est sa cohérence interne[25]. 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[26] 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 particulier des cas litigieux. Parallèlement aux corpus annotés, se développent, pour chaque schéma d'annotation, des guides

[25] C. Muller (1973, p. 10) le disait déjà voici longtemps, en particulier pour la segmentation et la lemmatisation.

[26] 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.

d'annotation (*guidelines*), qui sont parfois plus justement dénommés des « recueils de jurisprudence » (*caselaw*s). 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).

7. CONTRAINTES ET CONDITIONS INSTITUTIONNELLES

7.1 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[27]. 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.

7.2 Problèmes juridiques

Peu de corpus sont dans le domaine public sans condition aucune[28] : 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 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[29]. Lors de la

[27] 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.

[28] À l'exception, notable, de ***Susanne***, déchargeable par ftp anonyme (Sampson, 1994, p. 187) : black.ox.ac.uk (ota/suzanne).

[29] Le rapport de N. Pujol (1993) ne donne pas l'ensemble des situations qui peuvent se

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[30] 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.

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.

[30] « 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).