Le texte qui suit est un extrait de la thèse de Bénédicte Pincemin. Références complètes :
BOMMIER-PINCEMIN Bénédicte (1999) – *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de Doctorat en Linguistique, Université Paris IV Sorbonne, 6 avril 1999, chapitre VII : "Caractérisation d'un texte dans un corpus : du quantitatif vers le qualitatif", § A "Définir un corpus", pp. 415-427.

1. DÉFINIR UN CORPUS

a) Les données

Le corpus se définit de fait comme l'objet concret auquel s'applique le traitement, qu'il s'agisse d'une étude qualitative ou quantitative.

> *corpus* : (ling.) ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique ; (l'exicométrie) ensemble de textes réunis à des fins de comparaison, servant de base à une étude quantitative. (Lebart, Salem 1988, § *Glossaire*)

Mais les données ont un nom trompeur : elles ne s'imposent pas, elles sont construites. Certes, il y a un existant, directement sous forme de textes électroniques par exemple, –et donc l'analyste n'a pas une totale liberté d'« inventer » ses données, il part d'une réalité–, mais il reste des décisions du type : faut-il considérer tout ce qui est disponible ou en extraire un sous-ensemble plus significatif et équilibré ; comment tirer parti du codage disponible, comment éventuellement l'adapter au traitement envisagé. Le rapport aux données tient d'un compromis : faire avec ce à quoi on a accès, mais faire au mieux avec cela.

> La définition des textes et [le cas échéant des] fragments [qui subdivisent chaque texte] devrait dépendre du but de l'étude ; mais souvent, le statisticien ne peut qu'accepter les données disponibles...
> (Benzécri & *al.* 1981, p. 137)

Les linguistiques de corpus

L'accès actuel à de vastes ensembles de textes sous forme électronique a été une condition décisive pour le développement d'un courant linguistique récent : la linguistique à base de corpus (Habert, Nazarenko, Salem 1997).

L'approche à base de corpus revendique d'abord son réalisme, car elle se fonde sur des textes réels, des données attestées : le corpus s'oppose ici aux exemples ad hoc forgés pour les besoins d'une théorie ou d'une étude.

Le corpus est généralement l'apanage d'une linguistique descriptive, qui l'observe pour reconstituer a posteriori des régularités. Une linguistique normative peine à l'exploiter, car le corpus « brut » n'obéit pas au jeu de règles érigées a priori, si élaboré soit-il. Du côté des outils informatiques, le corpus appelle des traitements robustes, des analyses partielles.

b) Référentiel effectif

Le corpus fournit à la fois des éléments à étudier, mais aussi l'environnement descriptif de ces éléments. Le corpus est un tout, un vaste ensemble, qui constitue à lui seul le cadre et le référentiel de l'analyse Il met en présence les éléments, il fait qu'ils sont aussi considérés dans leur interrelation globale. Les éléments prennent alors une valeur relative par rapport au corpus : affinités et associations, fréquence ou rareté, banalité ou spécificité, etc.

Le cadre fixé par le corpus, souvent celui d'une application et d'une pratique, devient un moyen de réduire et d'ajuster l'appareil descriptif, grâce à un opportunisme efficace. On reprend et on adapte les ressources traditionnelles : ontologie et dictionnaire (limités au domaine), scripts (juste ceux associés aux situations envisageables dans la pratique concernée), lois de structuration du texte (sur la base de la forme conventionnelle du genre). Certains sombres problèmes des Traitements Automatiques des Langues trouvent soudain une issue : l'ambiguïté s'estompe, car dans un domaine fixé la langue prend un tour univoque ; l'implicite est dévoilé, puisque le corpus est ancré dans un cadre stéréotypé donné ; la granularité (ou niveau de détail) de la description trouve une juste mesure, en fonction de la définition du corpus et de l'application envisagée. (Pincemin, Assadi, Lemesle 1996, §7.1) (Péry-Woodley 1995, §3)

2. Le corpus : un ensemble de textes ?

a) Tout ensemble de textes n'est pas un corpus : propriétés recherchées

Le corpus ne se laisse pas uniquement définir formellement, comme un ensemble de texte ou une suite de caractères alphanumériques. Il vérifie trois types de conditions : des conditions de signifiance, des conditions d'acceptabilité, et des conditions d'exploitabilité.

- *Conditions de signifiance* : Un corpus est constitué en vue d'une étude déterminée (*pertinence*), portant sur un objet particulier, une réalité telle qu'elle est perçue sous un certain angle de vue (et non sur plusieurs thèmes ou facettes indépendants, simultanément) (*cohérence*).
- *Conditions d'acceptabilité* : Le corpus doit apporter une représentation fidèle (*représentativité*), sans être parasité par des contraintes externes (*régularité*). Il doit avoir une ampleur et un niveau de détail adaptés au degré de finesse et à la richesse attendue en résultat de l'analyse (*complétude*).
- *Conditions d'exploitabilité* : Les textes qui forment le corpus doivent être commensurables (*homogénéité*). Le corpus doit apporter suffisamment d'éléments pour pouvoir repérer des comportements significatifs (au sens statistique du terme) (*volume*).

Chacune de ces conditions demande à être commentée, à partir des éclairages complémentaires, et assez remarquablement convergents, issus des différentes disciplines qui utilisent les corpus (statistiques lexicales et lexicométrie, analyse de contenu en psycho-sociologie, linguistique structurale, etc.).

Pertinence

Le corpus prend sens par rapport à un objectif d'analyse. Cela n'est pas sans incidence sur la question de sa réutilisabilité : à quelles conditions ce qui a été rassemblé pour servir un objectif peut être recyclé pour en servir un autre ? Une partie de la réponse se trouve dans l'explicitation des choix et conditions de recueil du corpus. D'autre part, ce n'est pas nécessairement le corpus tel quel qui est repris : le corpus original sert de source pour construire un autre corpus, dans le respect du nouveau contexte d'analyse.

Règle de pertinence : Les documents retenus doivent être adéquats comme source d'information

pour correspondre à l'objectif qui suscite l'analyse. (Bardin 1977, §III.I.1, p. 128)

Cohérence

L'analyse du corpus mène à une représentation synthétique, qui doit donc, pour être claire et expressive, pouvoir être comprise comme la représentation d'une entité, avec ses articulations internes et non comme la juxtaposition de plusieurs réalités indépendantes. C'est par le même geste, que l'on se donne un corpus, et que l'on s'isole de toutes les problématiques générales ou étrangères.

Le caractère idiolectal des textes individuels ne nous permet pas d'oublier l'aspect éminemment social de la communication humaine. Il faut donc élargir le problème en posant comme principe qu'un certain nombre de textes individuels, à condition qu'ils soient choisis d'après des critères non linguistiques garantissant leur homogénéité, peuvent être constitués en corpus et que ce corpus pourra être considéré comme suffisamment isotope.

[...] ce qui permet [par exemple] de réunir une cinquantaine de réponses individuelles en corpus collectif, c'est un ensemble de caractères communs aux testés : leur appartenance à la même communauté linguistique, à la même classe d'âge ; c'est aussi le même niveau culturel, la même

« situation de testés ».

(Greimas 1966, §VI.3, pp. 93-94)

Règle d'homogénéité : les documents retenus doivent être homogènes, c'est-à-dire obéir à des critères de choix précis et ne pas présenter trop de singularité en dehors de ces critères de choix.

Par exemple, des entretiens d'enquête, effectués sur un thème donné, doivent : être tous concernés par ce thème, avoir été obtenus par des techniques identiques, être le fait d'individus comparables. Cette règle est surtout

utilisée lorsqu'on désire obtenir des résultats globaux ou comparer les résultats individuels entre eux. (Bardin 1977, §III.I.1, p. 128)

Lorsque nous utilisons [le terme corpus], nous sous-entendons 'corpus de documents homogènes', à savoir un ensemble de documents qui ne soit pas hétéroclite. Il ne s'agit pas de considérer n'importe quel ensemble de documents sans aucun rapport les uns avec les autres. Par exemple, un ensemble de brevets relatifs aux céramiques, un ensemble de publications mondiales sur l'intelligence artificielle constituent pour nous, des corpus homogènes. Les traitements que nous exposerons par la suite sont envisagés sur de tels corpus. (Chartron 1988, §II.1, p. 16)

Le choix d'un corpus présuppose... que ce corpus constitue bien un objet d'étude ; c'est-à-dire, que l'analyste le perçoive comme une entité ou un objet dans l'univers référentiel qui l'intéresse. En définitive, même si ce n'est que de manière implicite, l'analyste fait des hypothèses sur les conditions d'existence de cet objet, sur ses lois de production, sur les paramètres qui le font reconnaître dans cet univers référentiel. (Reinert 1990, §1.2, p. 27)

Représentativité

Les statisticiens soulignent bien que définir un échantillon est une opération complexe, pour assurer que l'extrait présente la même configuration des observables. La réalité à décrire présente un certain équilibre, une certaine composition, que le corpus doit d'efforcer de refléter.

Règle de représentativité : On peut, lorsque le matériel s'y prête, effectuer l'analyse sur échantillon. L'échantillonnage est dit rigoureux si l'échantillon est une partie représentative de l'univers de départ. Dans ce cas les résultats obtenus sur échantillon seront généralisables à tout l'ensemble.

Pour échantillonner il faut pouvoir repérer la distribution des caractères des éléments de l'échantillon. Un univers hétérogène demande un échantillon plus important qu'un univers homogène. [...] Comme pour un sondage, l'échantillonnage peut se faire au hasard, ou par quotas (les fréquences des caractéristiques de la population étant connues, on les reprend dans des populations réduites pour l'échantillon).

(Bardin 1977, §III.I.1, p. 127)

Pour la linguistique, ce qui autorise des études sur des corpus toujours limités, c'est la nature redondante de la langue et la clôture des unités textuelles.

Le corpus n'est [...] jamais que partiel, et ce serait renoncer à la description que de chercher à assimiler, sans plus, l'idée de sa représentativité à celle de la totalité de la manifestation. Ce qui permet de soutenir que le corpus, tout en restant partiel, peut être représentatif, ce sont les traits fondamentaux du fonctionnement du discours retenus sous les noms de redondance et de clôture. Nous avons vu que toute manifestation est itérative, que le discours tend très vite à se fermer sur luimême : autrement dit, la manière d'être du discours porte en elle-même les conditions de sa représentativité. (Greimas 1966, §IX.1.b, p. 143)

Quand l'étude vise à décrire la langue ou le fonctionnement des textes « en général », la condition de représentativité semble devoir se traduire par une recherche de diversité maximale. Autrement dit, dans l'idéal, tous les cas de figure existants doivent être présents dans le corpus. Deux tactiques sont observables : la course à la quantité d'une part (engranger le maximum de données, le poids total devant être garant de la richesse amassée), la construction raisonnée d'autre part (se donner une grille quadrillant la réalité, et s'en servir pour rassembler méthodiquement des textes correspondant à tous les aspects recensés). La première tactique, dont la devise est « more data is better data » (Péry-Woodley 1995, §2.3.1), est manifestement grossière, mais souvent elle est justifiée (en partie) par les difficultés profondes auxquelles se heurte de plein fouet la seconde tactique : quel modèle adopter pour organiser la sélection des textes, qui ne porte pas sa part d'a priori réducteurs ? Plus

gravement, la problématique elle-même apparaît utopique irréaliste : il n'y a pas de langue générale, ou standard, ou moyenne ; et les textes sont tous pris dans des pratiques qui les contextualisent1.

La recherche de corpus équilibrés semble bien constituer une impasse : la notion d'équilibre s'apparente à celle de « langue générale », et elle paraît tout aussi insaisissable. Elle suppose également une recherche irréaliste d'exhaustivité : le corpus équilibré est sans doute celui qui a « de tout un peu », mais encore faudrait-il savoir ce qu'est « tout », c'est-à-dire quelles sont les classes à représenter, –ce qui nécessite un modèle complet de la variation –, et avoir accès à des textes les représentant. (Péry-Woodley 1995, §2.3.2, p. 218)

Admettre la relativité et la part de choix qu'il y a dans la constitution de tout corpus, c'est également reconnaître le caractère décisif de l'établissement du corpus. En particulier, bien souvent le corpus (ou une de ses parties) est utilisé comme référentiel (puisqu'il est représentatif de la réalité à décrire) et il conditionne tous les résultats de l'analyse.

Le choix d'une norme endogène au corpus, le tout comme étalon des parties, est justifié par le fait maintenant bien établi qu'une forme [i.e. une unité], quelle qu'elle soit, n'a pas de fréquence en langue. (Note : Certains auteurs, contre toute évidence, affirment le contraire et invoquent des probabilités de langue. En revanche, nous sommes bien conscients du fait que l'usage d'une norme intrinsèque confère à l'élaboration du corpus une écrasante responsabilité.) (Lafon 1980, p. 137)

Régularité

La régularité correspond au fait que l'on explicite des principes pour définir le corpus, sans se permettre d'exceptions qui introduiraient des écarts locaux (manques, excès, éléments étrangers).

Règle de l'exhaustivité : une fois défini le champ du corpus (entretiens d'une enquête, réponses à un questionnaire, éditoriaux d'un quotidien de Paris entre telle et telle date, émissions de télévision concernant tel sujet, etc.), il faut prendre en compte tous les éléments de celui-ci. Autrement dit, il n'y a pas lieu de laisser un élément pour une raison quelconque (difficulté d'accès, impression de non-intérêt) non justifiable sur le plan de la rigueur. Cette règle est complétée par la règle de nonsélectivité.

Par exemple, on réunit un matériel d'analyse des publicités pour automobiles parues dans la presse pendant une année. Toute annonce publicitaire répondant à ces critères doit être recensée.

(Bardin 1977, §III.I.1, p. 127)

[Exigence d']exhaustivité : les ensembles [des individus et des variables] représentent un inventaire complet d'un domaine réel dont le cadre n'est guère discutable. (Benzécri & al. 1973b,

§A.2.1.3, p. 21)

Complétude

Le corpus doit avoir un niveau de détail adapté aux besoins de l'analyse : les adaptations nécessaires peuvent être soit de l'enrichir et de l'affiner, soit d'ajuster, par réduction, le niveau de discrétisation de la réalité à représenter réalisée à partir des données.

L'exhaustivité du corpus est [...] à concevoir comme l'adéquation du modèle à construire à la

[1] Une voie envisagée a donc été de s'appuyer sur une description systématique des situations de communication et de production des discours. On se donne un ensemble de paramètres, tels que : la communication directe (interlocution) ou différée, l'adresse à un public/lectorat collectif ou non, le caractère formel de l'échange, etc. C'est la méthode adoptée dans (Bronckart & *al.* 1985). Douglas Biber (Biber 1988) recule d'un cran le caractère nécessairement subjectif d'une telle grille, en se fondant non pas directement sur les pratiques de communication (et donc les genres), mais en partant d'un ensemble de caractéristiques linguistiques (essentiellement morpho-syntaxiques) pressenties comme liées à la diversité des genres. L'étude dépend donc toujours, mais cette fois-ci indirectement, d'une certaine perception que l'on a des genres. Même si la statistique (analyse factorielle) a un pouvoir certain de généralisation (gommage d'éléments non pertinents, interpolation à partir d'un nombre limité d'éléments, caractère suggestif des représentations), les résultats de Douglas Biber doivent être compris comme relatifs aux choix initiaux (textes utilisés pour l'étude, choix des traits morphosyntaxiques représentatifs).

totalité de ses éléments implicitement contenus dans le corpus. (Greimas 1966, §IX.1.b, p. 143)

exhaustivité : l'exhaustivité des données (qui assure à l'analyse une base intrinsèque [...]) peut,

conformément au principe d'équivalence distributionnelle, être assurée par une partition [...], ou [par le] choix d'un échantillon fini (éventuellement stratifié [...]) sur un espace potentiel continu (Benzécri & al. 1973, § Indice systématique)

Homogénéité

Sachant l'objectif de l'analyse, et les dimensions de variation que l'on veut étudier, le corpus doit être aussi homogène que possible pour ses autres caractéristiques.

[Exigence d']homogénéité : toutes les grandeurs recensées [...] sont des quantités de même

nature. (Benzécri & al. 1973b, §A.2.1.3, p. 21) homogénéité : pour définir objectivement le tableau des données étudiées [...], on vise à l'homogénéité des variables : ce qui permet l'adoption d'une unité de mesure unique [...] ; l'homogénéité est autorisée par l'hypothèse du nexus, [à savoir celle de l'] interrelation de tous les caractères d'un vivant (Benzécri & al. 1973, § Indice systématique)

Volume

Les procédés d'analyse visent à saisir et décrire des régularités qui structurent le corpus. Une certaine redondance est nécessaire pour que puissent émerger et être repérés des aspects caractéristiques et informatifs.

Le logiciel ALCESTE est un outil d'aide à l'interprétation d'un corpus textuel : entretiens, réponses à une question ouverte, textes littéraires, en fait tout document écrit à l'aide de l'alphabet latin, des dix chiffres et des signes usuels de ponctuation pourvu qu'il présente une certaine homogénéité et un volume minimum. [...]

Il y a toutefois deux conditions pour obtenir un résultat signifiant : la première est que le corpus présente une certaine cohérence thématique [cf. condition d'homogénéité]. C'est le cas (en général !) des réponses à une question ouverte, de textes littéraires, de recueils d'articles sur un sujet, etc... A contrario on ne peut pas espérer une indication de contenu pour un patchwork de fragments disparates, aussi intéressants soient-ils isolément...

La seconde est que le document soit suffisamment volumineux pour que l'élément statistique entre en ligne de compte. C'est du reste l'intérêt d'ALCESTE de donner très rapidement une vision globale sur une documentation volumineuse qui serait autrement très longue à dépouiller. (Reinert, Piat 1995, cahier 1, §0, p.3)

La condition de volume est importante pour des analyses statistiques, pour que celles-ci puissent être considérées significatives. En revanche, présenter la recherche de volume essentiellement comme un moyen d'obtenir une bonne représentativité 'générale' (Church & Mercer 1993) est déplacé : le volume et la représentativité sont des caractéristiques à part entière, complémentaires.

Dans le cas d'une exploitation manuelle, c'est-à-dire sans l'outil informatique, on s'inquiétera à l'inverse de la maniabilité du corpus (Garcia-Debanc 1989, p. 44).