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.

A. DÉFINIR UN CORPUS

1. Une question qui resurgit dans le contexte du calcul

Le corpus est nécessité et orienté par le traitement : c'est bien le préliminaire aux calculs, et c'est sous cet angle qu'il est considéré dans ce chapitre.

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 ; (lexicométrie) 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 lui-même : 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 contextualisent[1].

> 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

[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).

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 non-sélectivité.
>
> 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 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).

b) Du texte, des textes

Certains travaux ne considèrent pas les unités que forment les textes, ils ne visent que le matériau linguistique, à savoir une seule des facettes du texte. Le corpus est alors un ensemble de données pour des études de la langue.

> Nous employons le mot *corpus* dans une acception restreinte empruntée à J. Sinclair [...] : « Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage. » (Habert, Nazarenko, Salem 1997, p. 11)

C'est particulièrement à Marie-Paule Péry Woodley que l'on doit d'avoir questionné le bien-fondé de corpus linguistiques mais non textuels, qui rassemblent *du* texte et non *des* textes. En effet, ce choix s'apparente à un appauvrissement systématique et injustifié : toute manifestation linguistique ressort d'une forme de textualité, au sens d'une unité de communication. Et toute suite linguistique reçoit une part de sa définition et de sa significativité de son contexte textuel (par le genre auquel se rattache le texte, ou encore par le jeu des zones de localité, du syntagme à l'étendue du texte entier). S'en tenir à des extraits, même larges, ne suffit pas à rendre compte du fonctionnement global qu'institue l'unité texte. De plus, le corpus devient alors un bloc monolithique et inerte, car les dimensions qui auraient permis sa redéfinition pour une autre étude ont été effacées.

> Notons l'absence d'article devant le mot *text* dans la phrase de K. Church et R. Mercer citée plus haut [...] : il s'agit d'analyser *du* texte et non *des* textes. Se pose ici la question de la pertinence de l'unité texte dans la constitution et l'analyse de corpus : un ou des textes par opposition à du texte. Un corpus se compose par définition de discours, de langue concrète [...], et c'est immanquablement sous la forme de textes –écrits ou parlés– que la langue se réalise en discours. Cette langue concrétisée en textes porte les marques des conditions de leur production, et des objectifs qui les ont motivés. A l'extrême, recueillir du texte, c'est ne se donner aucun moyen d'échantillonnage, c'est soumettre ensuite à l'analyse un objet dont l'hétérogénéité est totalement opaque, c'est enfin se priver de toute possibilité de prise en compte de la structure textuelle. Les conséquences d'un tel choix se situeraient donc dans l'immédiat sur le plan de la qualité du travail possible sur un tel corpus, et au delà sur le plan de sa réutilisabilité. (Péry-Woodley 1995, §2.3.3, pp. 218-219)

3. Constitution : une typologie des corpus en présence

a) Emboîtements

En prenant le mot corpus dans son sens le plus large, il s'avère que l'analyste n'a pas affaire à un corpus –un ensemble de textes–, mais à une série de corpus[2], qui ont chacun leur rôle.

- *Le corpus existant (ou corpus latent)* : l'ensemble des textes auxquels on peut avoir accès, dont on peut disposer. C'est généralement une masse « informe », non systématique, mal défini, aux contours incertains. Il est difficile d'en avoir une vue globale. Cet existant dépend de conditions étrangères à l'étude, qui ne sont pas toutes connues ni maîtrisées.
- *Le corpus de référence* : il est composé à partir du corpus existant, en adéquation avec l'objectif de travail ; il est clairement défini et équilibré. C'est lui qui fournit l'univers le plus large dans lequel chaque élément trouve sa valeur. Il constitue l'univers et fixe le point de vue de l'étude. Il représente le fond sur lequel on veut profiler les textes étudiés. Autrement dit, il matérialise le contexte, actif et virtuel, et acquiert là son statut linguistique d'unité de rang supérieur –la linguistique ne s'arrêtant ni à la phrase, ni même au texte (Rastier 1998, §III.2).

[2] On pourrait aussi préférer réserver le terme *corpus* pour les ensembles de textes rassemblés pour l'analyse et qui en fournissent le contexte (*corpus de référence*, *corpus d'étude*). Le *fonds documentaire* désignerait alors l'ensemble des textes à disposition (de préférence à *corpus existant*) ; quant au *corpus distingué*, il correspond habituellement à ce que l'on appelle *sous-corpus*.
Le choix ici de décliner quatre *corpus*, malgré leurs différences profondes, se justifie par l'intention de souligner les usages contrastés du mot *corpus*.

- *Le corpus d'étude* : c'est l'ensemble des textes sur lesquels porte effectivement l'analyse, pour lesquels on attend des enseignements, des résultats. Le corpus d'étude n'est pas nécessairement une partie du corpus de référence, mais le corpus de référence doit pouvoir être considéré comme représentatif du corpus d'étude, pour l'aspect dont on veut rendre compte. Paradoxalement, le corpus d'étude peut être plus volumineux que le corpus de référence : ce qui est définitoire, ce n'est donc ni un rapport d'inclusion, ni un rapport de taille, mais la spécificité des rôles de chacun.
- *Le corpus distingué* : c'est un groupe de textes du corpus d'étude que l'on veut caractériser dans leur cohésion d'ensemble, par rapport au reste du corpus d'étude.

Exemples illustratifs, d'après des travaux actuels :

	corpus existant	**corpus de référence**	**corpus d'étude**	**corpus distingué**
Etude d'Etienne Brunet (Brunet 1995)	la base Frantext de l'INaLF	350 romans entre 1830 et 1970	phrases de ces romans comportant au moins une des 165 unités lexicales retenues pour définir la thématique du sentiment	les éléments retenus dans les romans d'un romancier
Constructio n des profils pour l'application DECID de diffusion ciblée	textes enregistrés dans la base SPHERE de la DER d'EDF, autres textes électroniques collectés de façon centralisée.	l'ensemble des textes d'Action, en version définitive, à partir de l'année 1990 jusqu'à l'année en cours.	les textes d'Action pour une année (le cas échéant, les textes en version provisoire pour l'année suivante).	les textes d'Action du corpus d'étude, dont le rédacteur (plus exactement le responsable) est rattaché à un Département donné.

Chaque choix est significatif, et joue un rôle pour la suite de l'analyse. Par exemple, Greimas montre l'incidence de ce qui est pour nous le corpus de référence :

> [Pour l'étude de l'univers de Bernanos,] la question pratique [...] est de savoir quelle signification il faut attribuer respectivement aux trois corpus possibles : le corpus ayant les dimensions d'un roman, le corpus de la totalité des écrits de Bernanos et, enfin, le corpus de tous les romans d'une société et d'une période historique données, et quelles corrélations structurelles on peut raisonnablement espérer retrouver entre les modèles qu'on pourra expliciter à partir de tels corpus.
>
> [...] d'une part, les corpus constitués par des romans-occurrences sont à considérer comme des inventaires de modèles implicites permettant la construction du genre « roman du XX^{ème} siècle » ; [...] d'autre part, les corpus faits de totalités représentatives de paroles individuelles constituent autant de manifestations pouvant servir à la construction d'un genre désigné sommairement comme « style de la personnalité » [...].
>
> Un roman-occurrence, le *Journal d'un curé de campagne*, [...] se trouve placé au croisement de deux axes, et [est] susceptible d'entrer simultanément dans deux corpus différents et d'être soumis à deux analyses ayant des visées divergentes. Pour ne prendre, à titre d'exemple, que l'analyse actancielle, on voit que les personnages de ce roman pourront être considérés comme les variables d'une structure actancielle romanesque propre à la littérature du XX^{ème} siècle, mais qu'ils participeront en même temps, comme des incarnations spécifiques, de la structure actancielle proprement bernanosienne.
>
> (Greimas 1966, §IX.1.f, pp. 148-149)

Une telle explicitation des articulations entre les différents ensembles de textes à considérer, chacun avec leur rôle dans l'étude, a déjà fait l'objet de réflexions, en Analyse du Discours par exemple :

> Nous introduisons [...] trois concepts complémentaires, ceux d'*univers discursif*, de *champ discursif* et d'*espace discursif*.
>
> On entendra par « univers discursif » l'ensemble des énoncés de tous types qui coexistent, ou plutôt interagissent, dans une conjoncture. Cet ensemble est nécessairement fini, mais irreprésentable, jamais pensable dans sa totalité par l'AD [Analyse du Discours]. Quand on utilise cette notion, c'est essentiellement pour y découper des « champs discursifs ».
>
> Le « champ discursif » est définissable comme un ensemble d'archives qui se trouvent en relation de concurrence, au sens large, et se délimitent donc pour une position énonciative dans une région donnée. Le découpage de tels champs doit découler d'hypothèses explicites et non d'une partition spontanée de l'univers discursif. Certes, la tradition a légué un certain nombre d'étiquettes (champs discursifs religieux, politique, littéraire, etc.), mais ce sont là des grilles extrêmement grossières, de peu d'intérêt pour l'AD, qui est contrainte à prendre en compte de multiples paramètres pour construire des champs pertinents.
>
> L'« espace discursif », enfin, délimite un sous-ensemble du champ discursif, lie au moins deux archives dont il est permis de penser qu'elles entretiennent des relations privilégiées, cruciales pour la compréhension des discours concernés. C'est donc une décision de l'analyste qui le définit, en fonction de ses objectifs de recherche. Si on découpe de tels sous-ensembles, ce n'est pas par simple

commodité (parce qu'il serait difficile d'appréhender un champ discursif dans sa totalité) mais aussi et surtout *parce qu'une archive donnée ne s'oppose pas de manière semblable à toutes celles qui partagent son champ* : certaines oppositions sont fondamentales, d'autres ne jouent pas directement un rôle essentiel dans la constitution et la préservation de l'archive considérée.

Aucun champ discursif n'est insulaire ; il existe une circulation intense d'une région à une autre de l'univers discursif, mais les voies qu'elle emprunte n'ont rien de stable ; selon les discours et les conjonctures concernés on aura affaire à des jeux d'échanges très différents. [...]

Cette étude des échanges entre champs débouche immédiatement sur la question de l'*efficacité* des discours, de leur aptitude à susciter l'adhésion d'un ensemble de sujets. Le réseau de renvois d'un champ à l'autre (qu'il s'agisse de citations explicites, de schèmes tacites, de captations,...) ne contribue pas peu à cette efficacité : confronté à un discours de tel champ, un sujet retrouve des éléments élaborés ailleurs qui, en intervenant subrepticement, créent un effet d'évidence. On assiste à une « métaphore », un transport généralisé d'un champ à l'autre (mais pas de n'importe quel champ à n'importe quel autre) sans qu'il soit possible de définir un lieu d'origine, un sens « propre » ; tout simplement parce que la question même de l'origine n'est pas pertinente ici.

(Maingueneau 1991, §4.3, pp. 158-159)

Il ne paraît pas abusif d'établir la correspondance suivante, même si la superposition des deux modèles n'est pas totale (par exemple, chez Maingueneau, l'espace discursif est inclus dans le champ discursif) :

Maingueneau :	univers discursif,	champ discursif,	espace discursif.
Pincemin :	corpus existant,	corpus de référence,	corpus d'étude.

Les travaux dans le domaine font également souvent état de sous-corpus. Ce qui est appelé *sous-corpus* est tantôt un corpus d'étude (pris comme une partie du corpus de référence), tantôt un corpus distingué (détaché du corpus d'étude). Le sous-corpus mérite une attention particulière, en ce qu'il est davantage qu'un corpus : non seulement, comme son nom l'indique, il entretient une relation privilégiée avec le corpus dont il est extrait ; mais aussi, sa nature est différente –il n'est pas toujours un ensemble de textes.

C'est cette question qui est examinée dans les paragraphes qui suivent, après une discussion sur le mode de contextualisation opérée par le corpus de référence.

b) L'intertexte : le corpus comme contexte et comme totalité

Déterminer le contexte est bien sûr un acte herméneutique majeur, puisque c'est décider ce qui est accessible, et même structurant, pour l'interprétation du texte, *en dernière instance.*

L'extension de corpus.

[...] le contexte sémantique d'un sémème n'a d'autres limites que celle du texte ; [...] les relations sémiques d'afférence peuvent excéder le contexte syntaxique, et relier des sémèmes en n'importe quel point du texte, avec un effet cumulatif ; cela est particulièrement clair avec les noms propres. Ce type d'extension repose toutefois sur une hypothèse forte : que le texte empirique est partout identique à lui-même dans la mesure où il mettrait partout en œuvre les mêmes types de systématicité.

On peut aussi étendre le corpus à l'ensemble de l'œuvre du même auteur. Selon une remarque (incidente) de Hjelmslev (1973, p.151), l'œuvre d'un auteur est la plus grande unité linguistique possible. Même dans le cas –privilégié– où l'on connaît l'auteur d'une série de textes, cette affirmation repose elle aussi sur une hypothèse forte : l'identité à soi de l'auteur –entendu comme idiolecte.

(Rastier 1987, §IX.4.2.1, p. 252)

Il s'agit d'un tout par rapport auquel se définit chaque texte, chaque élément du corpus étudié ; or, la linguistique nous avertit de trois totalités illusoires :

il faut abandonner trois totalités romantiques, séduisantes, mais infondées, sinon dans une ontologie : *(i)* Celle du texte [isolé] [...]. La notion de « clôture textuelle » chez les contemporains doit beaucoup à cet unitarisme romantique [...]. *(ii)* Celle de l'œuvre, à laquelle répond la notion de style individuel [...]. *(iii)* Celle de l'Intertexte, qui dérive de la notion schlegelienne de totalité littéraire. [...] Il n'est même pas exclu qu'aujourd'hui l'Hypertexte soit le dernier avatar de la Totalité romantique des textes. (Rastier 1998, §III.2, pp. 107-108)

Autrement dit, le texte ou l'œuvre sont des unités, que l'on peut étudier en tant que telles, mais non pas des totalités « définitives ». Si par exemple on étudie les romans d'un auteur isolément, il faut avoir conscience que l'on fait abstraction d'une dimension significative, la « profondeur » qui les met en relief dans un contexte intertextuel, comme l'appartenance à un genre. En revanche, ces unités sont d'excellents candidats au *corpus distingué*, et peuvent concourir, par un cheminement inverse d'extension (vs de focalisation), à la définition du corpus de référence.

A partir d'un texte [note : Nous convenons que le texte permet de recruter son intertexte, cf. (Rastier 1989, § 2)], l'intertexte est ce par quoi l'on accède par l'ensemble des références (ou allusions) et plus généralement par l'ensemble des connexions opérées par la lecture et qu'on peut

appeler l'*anagnose* [, selon la définition de Ioannis Kanellos et Théodore Thlivitis]. (Rastier 1998, §III.2, p. 108)

> Un texte n'est [pas] interprété « hors-contexte » mais au sein d'un *univers de textes*, que nous appelons *anagnose* et qui porte la trace d'une intention interprétative. [...]
>
> Un texte peut [...] appartenir à plusieurs intertextes, instanciant ainsi à sa mesure, différents points de vue. [...] l'intertexte constitue une sélection du lecteur, effectuée selon ses propres objectifs interprétatifs, et qui sert à « soutenir » les relations sémantiques qu'il désire mettre en avant.
>
> (Thlivitis 1998, §1.3 & 2.1.3, pp. 29 & 41)

Le *genre* est, lui, un candidat au corpus de référence, ou du moins un paramètre important de sa construction. Il rallie en effet le texte à une situation dans la réalité, dans les pratiques de rédaction et de lecture.

> Dans la problématique du texte, le contexte, contrôlé par le texte, se décline en zones de localité. Les éléments pertinents de la situation sont requis par l'analyse du texte : tout texte, par son genre, se situe dans une pratique. Le genre est ce qui permet de relier le contexte et la situation, car il est à la fois un principe organisateur du texte et un mode sémiotique de la pratique en cours. [...]
>
> Le texte semble certes en linguistique une unité maximale. Mais un point de vue plus philologique engage à considérer que l'ensemble des textes relevant d'un même genre (et d'une même langue) constitue un « bon » corpus au sein duquel il est possible de caractériser et d'analyser un texte. [note : le corpus est la seule objectivation possible (philologique) de l'intertexte, qui sinon demeure une notion des plus vagues.]
>
> (Rastier 1998, §III.2, p. 107)

Il faut toutefois garder leur autonomie aux deux concepts de *corpus* et de *genre*.

Les genres sont déterminés par les pratiques sociales. Ils sont reconnus et décrits par la linguistique, car c'est une réalité intertextuelle, par laquelle peuvent s'expliquer certaines affinités et certaines régularités entre des textes. Mais, notamment pour les besoins du codage et de la structuration des grandes bases textuelles, le genre court toujours le risque d'une définition théorique figée. Celui qui veut étiqueter et classer les textes d'un corpus par genre risque de voir la délimitation de ses « paquets » se dérober. Où commence et où finit le genre ? Parions que les discussions soient encore pour longtemps ouvertes.

Le corpus, lui, relève d'un point de vue, contingent, –parmi une multiplicité d'autres points de vue possibles[3]–, donnant un cadre à la constitution d'un objet. C'est un paramètre méthodologique, car l'étude veut que l'on se donne un domaine d'action, relatif à une recherche particulière[4]. Le corpus s'inscrirait davantage du côté de l'herméneutique que de celui de la linguistique. Il est défini par un objectif de lecture, d'analyse. Toute étude (de statistique textuelle, de texte) tôt ou tard, explicitement ou implicitement, se donne un corpus, dans lequel elle va piocher, calculer, contraster. Le corpus a été fixé, la manière dont cela a été fait peut être discutée, argumentée, mais de toutes façons la définition du corpus est effective.

En somme, un ensemble de textes relevant d'un même genre peuvent constituer un corpus intéressant et fructueux. Pour autant, ce n'est pas le seul corpus valide. Le genre est un facteur qui contribue à l'homogénéité du corpus, mais d'autres modes de cohésion sont possible.

c) Le sous-corpus est encore des textes

Eléments factuels

Un (sous-)corpus peut être constitué comme la réunion des textes qui ont un caractère factuel commun. Dans les bases bibliographiques, les caractères factuels sont prévus par des champs. On considère ainsi les œuvres de :

- un auteur,
- un genre,

---

[3] Sur la multiplicité des corpus / intertextes, voir notamment (Thlivitis 1998) : §1.2.2, p. 22 sq. (intertexte centré texte, ou centré auteur et plutôt descriptif, ou centré lecteur et plutôt productif) ; §2.1.3, p. 39 (incidence de l'intertexte sur ce que l'on perçoit dans la lecture d'un texte), et p. 41 (l'intertexte comme point de vue).

[4] Le caractère subjectif et singulier d'un corpus peut être relativisé, dans la mesure où il conduit à une exploitation et à des résultats s'inscrivant dans un cadre d'analyse commun et réutilisables. Voir par exemple l'effort de (Thlivitis 1998) dans ce sens, pour la réutilisation d''interprétations' (classes sémantiques) basées sur un corpus :

« la méthodologie de la Sémantique Interprétative Intertextuelle [...] oblige à l'explicitation des *sources sémiques* à l'origine de la constitution d'une classe sémantique. Le lecteur est donc indirectement incité à réutiliser les interprétations existantes, en y apportant sa propre interprétation. De cette manière, nous proposons le dépassement, à la fois, de la volonté d'atteindre une juste *objectivité descriptive* et de la liberté d'une *subjectivité descriptive* totale en les remplaçant par un *consensus inter-subjectif*, issu de l'interaction *multi-utilisateur* avec un *espace commun d'analyses* et soutenu par une méthode de travail interprétatif qui incite à la *consultation* et à la *réutilisation* récurrentes. » (Thlivitis 1998, §1.2.2.2, p. 28)

• une période,

ces critères pouvant être croisés (œuvres d'un genre à une époque, œuvres d'un auteur dans un genre, etc.)

On focalise ainsi explicitement l'étude, évitant des mélanges et des hétérogénéités qui brouilleraient l'analyse. En termes de normes linguistiques, différents degrés de régularités sont ainsi observables : la période reflète un *dialecte*, le genre un *sociolecte*, l'auteur un *idiolecte*.

> Dans l'analyse thématique, [le corpus] doit être restreint à bon escient pour pouvoir caractériser la spécificité des discours et des genres : les thèmes du roman ne sont pas ceux de l'essai ni du poème. Ainsi, en dépouillant un corpus trop étendu qui mêlait des romans et des essais dans la période 1830-1870, nous nous sommes aperçu que les sentiments du roman n'étaient pas ceux de l'essai. Par exemple, le sentiment de fraternité, récurrent dans les ouvrages de Leroux, et celui d'équité chez Proudhon, n'ont pas été relevés dans les romans, à l'exception confirmatrice des Misérables, qui alterne des chapitres romanesques et d'autres qui relèvent du genre de l'essai.
>
> A supposer même que le mot se rencontre dans des genres différents, rien n'assure qu'il se rapporte aux mêmes thèmes : *amour* se rencontrera certes en poésie et dans le roman, mais le thème de l'Amour diffère pourtant avec ces genres. Il n'a pas la même molécule sémique, ni les mêmes lexicalisations, ni les mêmes antonymes ».
>
> (Rastier 1995a, §II.1, p. 235)

Ce mode de construction est plus complexe qu'il n'y paraît. Rassembler l'œuvre d'un auteur : jusqu'où aller ? faut-il considérer telle œuvre secondaire, tel brouillon, tel écrit atypique,... ? Si le critère est le genre : comme il n'y a pas de consensus sur la liste et le contenu des genres, est-ce que par exemple ce qui est déclaré comme roman ou comme poésie correspond à ce que l'on veut étudier ? (le roman par lettres est-il un roman, le poème en prose fait-il partie de la poésie ?). La période suppose elle aussi un travail de définition : notre culture raisonne par siècles ; mais ce découpage n'est-il pas arbitraire et brutal ?

> L'analyse qui cherche à expliciter les transformations diachroniques des structures ne doit pas utiliser le découpage du texte [corpus] en tranches, correspondant aux « pleines » réalisations des structures, mais opérer, au contraire, une division du texte en séquences superposées, comprenant chaque fois, des deux côtés de la zone franche, deux zones d'enchevêtrement où les structures survivantes coexistent avec les structures de remplacement nouvellement élaborées (Greimas 1966, §IX.1.g, p. 151)

Morceaux choisis : les textes dans les textes

On peut choisir d'étudier les divisions structurelles d'un texte comme autant de « textes » autonomes : chapitres (pour le roman), actes (pour le théâtre), etc. Considérer que l'on obtient encore des textes, c'est envisager le caractère fractal de la textualité : en délimitant un passage, l'auteur lui compose un début, une fin, une cohérence propre ; et chaque texte semble receler en puissance une multitude d'autres textes plus petits.

Un passage peut être délibérément isolé, lu et étudié pour lui-même. La décision de le définir est un acte herméneutique, qui décide de lui reconnaître et de lui assigner les propriétés d'un texte. Les recueils de « textes » et les anthologies sont bien des collections de tels passages.

> Nous dirons qu'un énoncé, un texte ou un corpus est un ensemble de mots ; un ensemble ordonné, obéissant aux lois d'un idiome, et dont la suite naturelle est porteuse d'une signification. On convient d'autre part de nommer *texte* tout énoncé ou toute succession dénoncés, tout discours ou fragment de discours, sans limitation d'étendue, provenant d'un même locuteur ou scripteur et présentant une certaine unité ; une collection définie de textes sera nommée *corpus.* (Muller 1977, p. 5, §1)

d) Le sous-corpus est du texte (réunion d'extraits)

Contextes d'un concept pôle

Le concept consiste en un mot ou un ensemble de mots : une thématique pressentie comme importante, ou encore les désignations d'un personnage principal (héros)[5]. Les contextes sont alors établis, par une condition de proximité plus ou moins élaborée : fenêtre (en nombre de mots), zone typographique (phrase délimitée par la ponctuation forte, paragraphe), construction syntaxique (par exemple adjectifs qui qualifient le nom pôle). Bien entendu, toute la suite de l'étude est orientée par ces choix initiaux : choix des pôles surtout, et choix du mode de détermination des contextes[6]. Le

[5] Pour une étude autour d'un personnage, voir par exemple (Dupuy 1993).

[6] Un des reproches adressés à une méthode d'Analyse du Discours, est de s'en tenir à un corpus de phrases extraites sélectionnées par des termes-pivot (mots pôles), sans y reconnaître un choix interprétatif majeur, qui relativise les résultats de l'étude à un point de vue :

« J.-M. Marandin, J. Guilhaumou et D. Maldidier, J.-J. Courtine, de manières convergentes, ont montré les limites d'une méthode *qui repose sur un savoir a priori*, celui qui préside à la sélection par le chercheur des termes-pivots : 'L'analyse répond à la question de l'analyste ; mais en présentant cette réponse comme structure

statut d'un pôle peut également varier, d'objet central et référentiel de l'étude, à prétexte provisoire ou amorçage.

> Si l'on veut préciser encore le rapport de l'analyse lexicale à l'analyse thématique, il faut préciser que le mot à partir duquel commence la recherche n'en est pas l'objet, à la différence d'un mot-vedette qui ferait l'objet d'une recherche lexicographique. On va certes chercher, en utilisant les moyens d'assistance informatisés, d'autres mots et expressions qui sont cooccurrents. Une fois interprétés, les cooccurrents pour lesquels on aura identifié une relation sémantique seront considérés comme des corrélats, c'est-à-dire comme des lexicalisations complémentaires de la même molécule sémique.
>
> Le réseau des corrélats relie les manifestations lexicales du thème. Mais il faut pouvoir discerner le(s) meilleur(s) point d'entrée(s) dans ce réseau. La « vedette » n'est alors qu'un mot d'entrée, choisi pour sa fréquence, et dans l'hypothèse qu'il présente une lexicalisation synthétique du thème que l'on cherche à décrire.
>
> (Rastier 1995a, §II.1, p. 236)

Les textes de l'ensemble des pages Web sélectionnées en réponse à une requête, soumise à un moteur de recherche sur Internet, est une forme de (sous-)corpus de ce type. Nos réserves viennent du fait que l'ensemble des pages indexées par le moteur est une réalité mouvante et mal définie, sans logique d'ensemble, et donc ne forment pas un corpus qui vérifierait les critères énoncés ci-avant. Néanmoins, cet ensemble de pages, contenant toutes un ou plusieurs mots fixés, est un ensemble délimité, motivé, présentant une relative homogénéité thématique : il peut devenir le lieu d'analyses ciblées[7], inenvisageables sur Internet dans son ensemble.

Si la procédure de sélection des contextes est assez ciblée, la démarche peut devenir itérative : les nouveaux éléments contextuels sélectionnés lors d'une passe deviennent les pôles pour la passe suivante. C'est alors le fait que les contextes soient très sélectifs qui doit assurer la convergence, à savoir qu'après un certain nombre d'itérations, il n'y a plus d'éléments nouveaux sélectionnés (sans pour autant avoir sélectionné tous les mots du corpus !).

Greimas adopte une telle démarche itérative pour son étude à partir de l'opposition *vie* vs *mort*, dans l'œuvre de Bernanos :

> La procédure d'extraction apparaît donc, dans son ensemble, comme une série d'opérations successives d'extraction, chaque inventaire de contextes extraits permettant l'extraction et la mise en inventaire d'autres contextes, et ainsi jusqu'à épuisement du corpus, c'est-à-dire jusqu'au moment où la dernière extraction (*n*), utilisant le dernier inventaire (*n* - 1), ne fera plus apparaître de nouvelles qualifications. Cela voudra dire que le corpus utilisé pour fournir par extraction les éléments de signification appartenant à l'isotopie de *vie* et *mort*, choisie au départ, est épuisé de façon exhaustive. (Greimas 1966, §XII.1.b, p. 224)

Autre sélection motivée

Il s'agit de contraster un extrait (pas nécessairement continu), représentant une partie identifiée, par rapport au reste, ou encore de diviser (répartir) les textes en composantes homogènes. C'est ainsi que l'on peut choisir d'opposer les passages en style direct à la narration, les textes des différents personnages d'une pièce de théâtre.[8]

Echantillonnage

L'opération consiste à démultiplier un corpus pour pouvoir l'étudier sous la forme de plusieurs échantillons, chacun étant *a priori* représentatif de l'ensemble, avec quelques variantes locales jugées mineures (en fait, qui ne sont pas l'objet premier de l'analyse). L'avantage recherché

---

de base d'un texte, l'analyste fait un passage à la limite où il confond son intérêt et ce qu'est le discours' [écrit J.-M. Marandin]. Choisir des termes-pivots, c'est définir les thèmes du discours ; or, dans la méthode des termes-pivots, ce n'est pas le texte qui permet de repérer ces thèmes, mais les présupposés de l'analyste » (Maingueneau 1991, §3.1, pp. 82-83)

Maingueneau poursuit en présentant une méthodologie de détermination de termes-pivots, qui s'appuie sur les constructions linguistiques :

« Courtine propose [...] de renverser ce problème de délimitation des thèmes du discours en posant la question suivante : '*Comment dans le discours lui-même et par le discours lui-même* un élément déterminé peut-il être caractérisé comme thème du discours ? (comment, c'est-à-dire : par la présence de quelles structures, sous quelle forme linguistique ?)'. [...] Cette option conduit naturellement Courtine à s'intéresser aux structures syntaxiques de la thématisation et, parmi celles-ci, tout particulièremet aux formules du type : 'C'est X que P', 'Ce que P c'est X', 'X c'est ce que P'. » (Maingueneau 1991, §3.1, p. 87)

[7] Voir par exemple Live Topics, de François Bourdoncle (Laboratoire de l'Ecole des Mines), qui opère sur les résultats de recherche d'Alta Vista.

[8] (Dupuy 1993, p. 261 sq.) prend la liberté de construire six « pseudo-textes », entre lesquels se répartissent les phrases de son corpus initial en fonction de leur « niveau dialogal » (est-elle prise en charge par le narrateur, par un personnage) et des temps de leur(s) verbe(s). La continuité linéaire des textes, et même les délimitations entre les différents textes du corpus (il s'agit d'un recueil de nouvelles) sont donc purement et simplement éliminées.

peut être non seulement de posséder plusieurs « images » d'un même texte, mais aussi d'avoir des unités textuelles à caractériser pas trop longues, ou de taille régulière. En effet, la taille des échantillons est une décision extérieure au corpus.

> Un texte, considéré comme un ensemble, peut aussi être traité comme formé de plusieurs sous-ensembles. On peut soit en considérer les *divisions naturelles*, soit y créer des *divisions artificielles*. [...] Dans un roman, on pourrait considérer comme un sous-ensemble toutes les répliques des personnages (discours direct [...]), d'autre part tout le reste, où l'auteur se présente comme locuteur. Il va sans dire que ces divisions fournissent des fragments d'étendue variable et inégale. Même les cinq actes d'une pièce classique ont rarement le même nombre de vers. Nous réservons aux sous-ensembles ainsi créés [en suivant des divisions naturelles] le nom de *fragments*.
>
> On peut au contraire diviser un texte en *tranches* d'égale étendue sans tenir compte des divisions naturelles. On peut même constituer ces tranches par des segments prélevés en divers endroits du texte. Ainsi, pour diviser le texte de *Phèdre* (1 654 vers) en 10 tranches égales, à 1 vers près, donc de 165 ou 166 vers chacune, on peut soit couper aux vers 165, 330, 496, etc., soit prendre pour une première tranche les vers 1, 11, 21,..., 1641, 1651 ; pour une seconde les vers 2, 12, 22,..., 1642, 1652, et ainsi de suite [...]. Par ce dernier moyen, les tranches se rapprochent des échantillons aléatoires.
>
> [...] La réunion d'un grand nombre de textes indexés constitue un corpus, à l'intérieur duquel on se propose d'étudier et de quantifier certains faits lexicaux, syntaxiques, etc. Le corpus comprend en général des divisions ou sous-corpus, qui la plupart du temps ont une unité propre (chronologique, stylistique, etc.) ; dans ce cas ils entrent dans la catégorie des fragments telle qu'elle a été définie ci-dessus ; s'ils ont été mesurés de façon à avoir la même étendue, ils se rapprochent des tranches, mais sans toutefois avoir été obtenus par une procédure aléatoire.
>
> (Muller 1973, §3, pp. 15-16)