Chapitre 2

Les unités de la statistique textuelle

La nécessité de comparer des textes sur des bases quantitatives se présente aux chercheurs dans des domaines scientifiques très divers. Dans chaque cas particulier, le recours aux méthodes quantitatives est motivé par des préoccupations différentes et les objectifs poursuivis souvent très distincts (études stylométriques comparées de textes dus à différents auteurs, typologies des réponses d'individus à une même question ouverte, recherche documentaire, etc.).

L'expérience du traitement lexicométrique d'ensembles textuels réunis à partir de problématiques différentes montre, cependant, que, moyennant une adaptation minime, un même ensemble de méthodes trouve des applications pertinentes dans de nombreuses études de caractère textuel. C'est à l'exposé de ces méthodes que seront consacrés les chapitres qui suivent.

2.1 Le choix des unités de décompte.

Segmentation, identification, lemmatisation, désambiguïsation.

La méthode statistique s'appuie sur des mesures et des comptages réalisés à partir des objets que l'on veut comparer. Décompter des unités, les additionner entre elles, cela signifie, d'un certain point de vue, les considérer, au moins le temps d'une expérience, comme des occurrences identiques d'un même type ou d'une *forme* plus générale. Pour soumettre une série d'objets à des comparaisons statistiques il faut donc, dans un premier temps, définir une série de liens systématiques entre des cas particuliers et des catégories plus générales.

Dans la pratique, l'application de ces principes généraux implique que soit définie une *norme* permettant d'isoler de la chaîne textuelle les différentes unités sur lesquelles porteront les dénombrements à venir. L'opération qui

permet de découper le texte en *unités minimales* (c'est à dire en unités que l'on ne décomposera pas plus avant) s'appelle la *segmentation* du texte. A cette phase, qui permet d'émietter le texte en unités distinctes succède une phase de regroupement des unités identiques : la phase d'*identification* des unités textuelles.

Pour un même texte, les différentes normes de dépouillement ne conduisent pas au mêmes décomptes. Pour chaque domaine de recherche particulier, elles ne présentent pas toutes le même degré de pertinence, ni les mêmes avantages (et inconvénients) quant à leur mise en oeuvre pratique.

On comprendra, par exemple, qu'un chercheur qui explore un ensemble d'articles rassemblés au sein d'une base de données ayant trait au domaine de la Chimie, exige de voir regroupés en une même unité le singulier du substantif *acide* et son pluriel *acides* afin de pouvoir interroger lors d'une même requête chacun des textes sur la présence ou l'absence de l'une ou l'autre des formes dans l'ensemble des textes qu'il étudie.

Dans le domaine de l'étude des textes politiques, au contraire, les chercheurs ont constaté que le singulier et le pluriel d'un même substantif renvoient souvent à des notions différentes, parfois en opposition (cf. par exemple l'opposition dans les textes récents de *défense de la liberté /défense des libertés* qui renvoie à des courants politiques opposés). On préférera souvent, dans ce second cas, indexer séparément les deux types d'unités qui seront étudiées simultanément.

Au-delà de ces considérations propres à chaque domaine, une fois définie la norme de segmentation, les méthodes de la *statistique textuelle* s'appliquent sans adaptation particulière aux décomptes réalisés à partir de chacune des normes, même si le rappel précis des contours de chaque unité textuelle, reste présent dans l'esprit du chercheur lors des phases suivantes de l'analyse.

Le problème de la définition des unités les plus aptes à servir de base aux analyses quantitatives a longtemps opposé les tenants du découpage en *formes graphiques* (directement prélevables à partir du texte stocké sur support magnétique) et les tenants de la lemmatisation qui jugent nécessaire de mettre en oeuvre des procédures d'identification plus élaborées, rassemblant l'ensemble des flexions d'une même unité de langue.

L'évolution simultanée des centres d'intérêt dans la recherche lexicométrique, des méthodes de la statistique textuelle mais aussi les avancées technologiques dans le domaine de la micro-informatique, permettent aujourd'hui de mieux cerner les spécificités et la complémentarité des deux approches, et d'entrevoir l'ensemble de ces problèmes sous un jour nouveau.

2.1.1 Le texte en machine

Sur toutes les machines permettant de saisir et de stocker du texte, on dispose désormais d'un système de caractères, qui compte en général une centaine d'éléments. Parmi ces caractères, certains correspondent aux lettres de l'alphabet : majuscules, minuscules, lettres diacrisées (munies d'accents, par exemple) propres à la langue que l'on traite ; d'autres servent à coder les chiffres ; d'autres encore permettent de coder des signes comme le pourcentage, le dollar etc. ; enfin, certains caractères servent à coder les divers signes de ponctuation usuels.

Les systèmes actuels individualisent un caractère particulier le "retour-chariot" qui permet de séparer des *paragraphes* (précisément définis comme l'ensemble des caractères situés entre deux retours-chariots).

Dans la pratique, un ensemble de normes typographiques unique se met progressivement en place pour réaliser l'opération que l'on appelait précédemment : l'*encodage* des textes. Ces normes subissent et subiront de plus en plus les effets des progrès technologiques dans le domaine la saisie des textes.

Aujourd'hui, lorsqu'ils ne sont pas directement composés sur clavier d'ordinateur, les textes peuvent être mis à la disposition des chercheurs par simple *scannage* (opération comportant la reconnaissance des différents caractères) de leur support papier. De ce fait, la masse des textes mis à la disposition des chercheurs pour pratiquer des études lexicométriques est désormais sans commune mesure avec celle dont disposait la communauté scientifique il y a tout juste vingt ans.

C'est circonstances renforcent le besoin de disposer d'outils méthodologiques relativement simples permettant d'inventorier, de comparer, d'analyser des informations textuelles toujours plus volumineuses.

2.1.2 Les dépouillements en formes graphiques

Le dépouillement en formes graphiques constitue un moyen particulièrement simple de constituer des unités textuelles à partir d'un corpus de textes. Suivant les objectifs de l'étude on accordera à cette approche un statut qui pourra varier : vérification de la saisie, première approche du vocabulaire, base des comparaisons statistiques à venir.

Pour réaliser une segmentation automatique du texte en occurrences de formes graphiques, il suffit de choisir parmi l'ensemble des caractères un

sous-ensemble que l'on désignera sous le nom d'ensemble des *caractères délimiteurs* (les autres caractères contenus dans la police seront de ce seul fait considérés comme caractères non-délimiteurs).

Une suite de caractères non-délimiteurs bornée à ses deux extrémités par des caractères délimiteurs est une *occurrence*. Deux suites identiques de caractères non-délimiteurs constituent deux occurrences d'une même *forme*. L'ensemble des formes d'un texte constitue son *vocabulaire*.

La segmentation ainsi définie permet de considérer le texte comme une suite d'occurrences séparées entre elles par un ou plusieurs caractères délimiteurs. Le nombre total des occurrences contenues dans un texte est sa *taille* ou sa *longueur*.

Cette approche suppose qu'à chacun des caractères du texte correspond un statut et un seul, c'est à dire que le texte à été débarrassé de certaines ambiguïtés de codage (par exemple : points de fin de phrase et points pouvant être présents à l'intérieur de sigles ou d'abréviations : S.N.C.F, etc.)[1].

2.1.3 Les dépouillements lemmatisés

Privilégiant le point de vue lexicographique, on peut, dans certaines situations, considérer qu'il est indispensable, avant tout traitement quantitatif sur un corpus de textes, de soumettre les unités graphiques issues de la segmentation automatique à une *lemmatisation*, c'est-à-dire de se donner des règles d'identification permettant de regrouper dans de mêmes unités les formes graphiques qui correspondent aux différentes flexions d'un même lemme.

Pour lemmatiser le vocabulaire d'un texte écrit en français, on ramène en général :

- les formes verbales à l'infinitif,
- les substantifs au singulier,
- les adjectifs au masculin singulier,
- les formes élidées à la forme sans élision.

Cette manière d'opérer, qui vise à permettre des décomptes sur des unités beaucoup plus soigneusement définies du point de vue de la "langue", peut

[1] Notons qu'un premier dépouillement en formes graphiques constitue souvent le moyen le plus sûr pour répertorier les problèmes de ce type qui peuvent subsister au sein d'un corpus de textes que l'on désire étudier. On prêtera également attention aux différences induites par l'utilisation des majuscules (début de phrase ou nom propre) aux traits d'unions, etc.

paraître séduisante au premier abord. Cependant, la pratique de la lemmatisation du vocabulaire d'un corpus rencontre inévitablement des problèmes dont la solution est parfois difficile[1].

En effet, s'il est relativement aisé de reconstituer, "en langue", l'infinitif d'un verbe à partir de sa forme conjuguée, le substantif singulier à partir du pluriel etc., la détermination systématique du lemme de rattachement pour chacune des formes graphiques qui composent un texte suppose, dans de nombreux cas, que soient levées préalablement certaines ambiguïtés.

Certaines de ces ambiguïtés résultent d'une homographie "fortuite" entre deux formes graphiques qui constituent des flexions de lemmes très nettement différenciés (par exemple : *avions* issu du verbe *avoir,* et *avions:* substantif masculin pluriel). Pour d'autres il s'agit de dérivations ayant acquis des acceptions différentes à partir d'une même souche étymologique (cf. les différents sens du mot *voile*, par exemple).

Dans certains cas il faut lever des ambiguïtés touchant à la fonction syntaxique de la forme, ce qui nécessite une analyse grammaticale de la phrase qui la contient.

Certaines de ces ambiguïtés d'ordre sémantique peuvent être levées par simple examen du contexte immédiat. D'autres nécessitent que l'on regarde plusieurs paragraphes, voire l'ensemble du texte[2].

Parfois enfin l'ambiguïté entre plusieurs sens d'une forme, plus ou moins entretenue par l'auteur, ne peut être levée qu'au prix d'un choix arbitraire.

Dans son *Initiation à la statistique linguistique*, Ch. Muller (1968) expose les difficultés liées à l'établissement d'une telle norme de dépouillement

> La norme devrait être acceptable à la fois pour le linguiste, pour ses auxiliaires, et pour le statisticien. Mais leurs exigences sont souvent contradictoires. L'analyse linguistique aboutit à des classements nuancés, qui comportent toujours des zones d'indétermination; la matière sur laquelle elle opère est éminemment continue, et il est rare qu'on puisse y tracer des limites nettes; elle exige la plupart du temps un examen attentif de l'entourage syntagmatique (contexte) et paradigmatique (lexique) avant de trancher. La statistique, dans toutes ses applications, ne va pas sans une certaine simplification des catégories; elle ne pourra entrer en action que quand le continu du langage à été rendu discontinu, ce qui est plus difficile

[1] Cf. par exemple Bolasco (1992) et les travaux de son laboratoire (Universita di Salerno).

[2] Dans certains analyseurs morpho-syntaxiques automatiques, l'ambiguïté peut être levée à partir d'une modélisation de certaines régularités statistiques (cf. par exemple : Bouchaffra et Rouault, 1992).

au niveau lexical qu'aux autres niveaux; elle perd de son efficacité quand on multiplie les distinctions ou quand on émiette les catégories; opérant sur des ensembles très vastes elle tolère difficilement une casuistique subtile qui s'arrête à analyser les faits isolés.

Dans la pratique, les tenants de la lemmatisation[1] s'appuient pour effectuer leurs dépouillements sur le découpage "en lemmes" opéré par un dictionnaire choisi au départ. Les décisions lexicographiques prises par ce dictionnaire serviront de références tout au long du dépouillement.

De fait, certaines études scindent en deux unités distinctes les formes contractées qui correspondent à une seule unité graphique du texte (*au -> à +le; aux -> à les*, etc.). D'autres regroupent en une même unité certaines unités graphiques qui correspondent à des locutions et dont la liste varie selon les études.

Pour des ensembles de textes peu volumineux, la pratique de la lemmatisation "manuelle" peut constituer une source de réflexions utiles portant à la fois sur le découpage lexicographique opéré "en langue" par les dictionnaires et sur les emplois particuliers attestés dans les corpus de textes considérés. Cependant, ces appréciations positives doivent être sensiblement nuancées lorsqu'on envisage des dépouillements portant sur des corpus de textes étendus.

L'examen des problèmes liés à la lemmatisation montre qu'il ne peut exister de méthode à la fois fiable et entièrement automatisable permettant de ramener à un lemme chacune des unités issues de la segmentation d'un texte en formes graphiques.

2.1.4 Les dépouillements à visée "sémantique"

On a regroupé sous cette rubrique différentes démarches utilisées dans des domaines n'entretenant parfois que peu de liens entre eux, mais pour lesquels le texte de départ subit un précodage substantiel. Le double but assigné à cette dernière opération est, à la fois d'*informer* le texte stocké en machine sur toute une série de traits sémantiques liés à chacune des unités dont il est composé, tout en utilisant, par ailleurs, des procédures définies de manière relativement formelle.

---

[1] On trouvera le point de vue des "lemmatiseurs" dans les différents travaux de Ch. Muller. On doit également à cet auteur une critique, relativement mesurée, du point de vue opposé, que l'on trouvera dans la préface *De la lemmatisation* qu'il a donné à l'ouvrage Lafon (1981). Le point de vue des "non-lemmatiseurs" est exposé dans les travaux du laboratoire de lexicométrie politique St.Cloud. Voir par exemple Geffroy et al. (1974) et la réponse de M. Tournier (1985a) à Ch. Muller.

Les "frames"

La représentation par "frames", proposée par Roger Schank (1975), se fixe pour objet "la représentation du sens de chaque phrase". Cette démarche constitue un exemple extrême d'une méthodologie qui tend à insérer le discours dans des catégories sémantiques définies a priori et de manière définitive. La représentation de Schank est censée prévoir à l'avance, dans ses grandes lignes du moins, toutes les questions qui peuvent se poser à propos d'une unité textuelle dans le cadre de ce qu'il appelle *la vie de tous les jours*. L'exemple qui suit donne une première idée du fonctionnement et du cadre conceptuel d'un tel mécanisme.[1]

La phrase : *Marie pleure* reçoit dans ce type de représentation le codage esquissé sur la figure 2.1

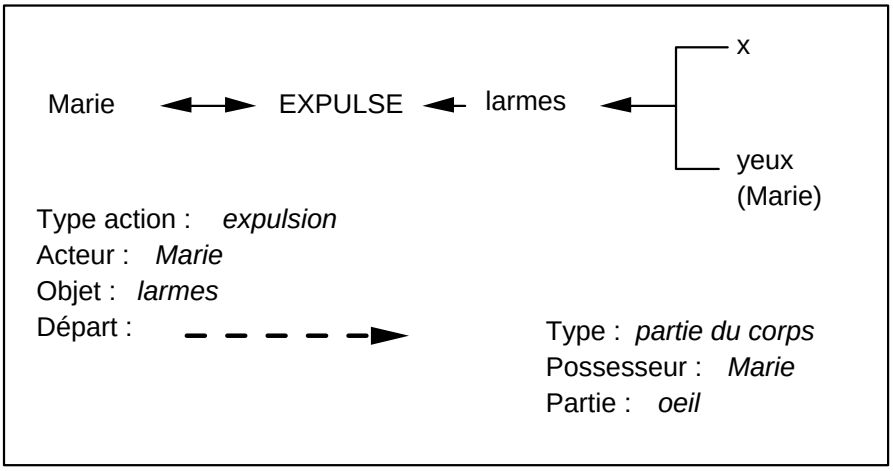

Figure 2.1

Représentation schématisée de la phrase "Marie pleure"

Diverses publications dans le domaine de l'intelligence artificielle témoignent de l'utilité de telles entreprises pour la mise en place de systèmes experts adaptés à un domaine de connaissance particulier.

Cependant les préoccupations et la pratique de l'analyse de textes dans nombreux domaines des sciences humaines nous laisse penser qu'un tel prédécoupage a priori de la réalité conduit à une mutilation considérable du matériau textuel soumis à l'analyse.

Analyse de contenu

[1] Cet exemple est reproduit dans l'ouvrage de Pitrat (1985).

Bien qu'elles fassent intervenir un codage beaucoup plus complexe du monde réel, les pratiques d'analyse textuelle utilisées au sein du courant que l'on a coutume d'appeler *analyse de contenu*[1] ne sont pas sans présenter des similitudes avec ce qui vient d'être évoqué plus haut. Dans ce cas encore, déjà évoqué au premier chapitre, les unités textuelles sont rassemblées, avant le recours aux comptages, dans des classes définies a priori, ou après une première lecture du texte, afin que les comptages portent directement sur les "contenus".

Ici encore les problèmes naissent de la grande latitude laissée en pratique à l'utilisateur dans la définition des catégories de comptages.

Familles morphologiques

M. Reinert a développé une méthodologie qui se fixe pour objet, de dresser une typologie des *unités de contexte* (séquences de textes de longueurs comparables, qui peuvent souvent coïncider avec les phrases) contenues dans un texte, fondée sur les associations réalisées par ces dernières parmi les unités appartenant à une même famille morphologique[2].

Dans une première phase entièrement automatisée, le logiciel *Alceste*, après avoir écarté de l'analyse toute une série de mots-outil contenus dans un glossaire, regroupe les formes susceptibles d'appartenir à une même famille morphologique, ignorant délibérément tout ce qui peut apparaître comme désinence finale. Lors du dépouillement de la traduction en français d'un roman de L. Durrell, l'algorithme a ainsi rassemblé dans une même classe notée *admir+* les formes graphiques :

admir+ : *admirable, admirait, admirateurs, admiration, admire, admirée, admiriez, admires*

Certaines classes construites par cet algorithme à partir des données de départ, résultent de coïncidences graphiques (dans le cas présent : *acide* et *acier*) et seront éliminées dans un second temps, lors de l'examen détaillé par le chercheur, des classes proposées par le programme.

Cette manière de procéder augmente considérablement le nombre des liens que l'on peut établir entre les différentes unités graphiques d'un même texte, circonstance particulièrement intéressante dans le cas de l'analyse de corpus de petite taille.

---

[1] Cf. par exemple Bardin (1977), Weber (1985).

[2] Cf. Reinert (1986, 1990). On trouvera une brève description du logiciel *Alceste* en annexe.

2.1.5 Très brève comparaison avec d'autres langues

Les problèmes qui touchent à l'utilisation de la forme graphique dans les études lexicométriques s'éclairent d'un jour particulier lorsqu'on compare le cas du français à celui d'autres langues. Les quelques exemples qui suivent n'ont d'autre prétention que de faire entrevoir au lecteur la variété des problèmes auxquels s'affronte la délimitation automatique de l'unité de décompte. Comme on le verra, ces différences entre langues concernent à la fois les particularités morpho-syntaxiques qui touchent en profondeur à la structure de chacune des langues concernées et la tradition orthographique qui leur est propre.

L'anglais regroupe, par exemple, par rapport au français, de nombreuses flexions d'une même forme verbale, sous une forme graphique unique : (*speak* versus : *parle, parles, parlons, parlez, parlent, etc.).* Il en va de même pour les flexions qui correspondent, en français, au genre de l'adjectif.

En espagnol, les pronoms personnels : *nos, se*, pronoms enclitiques, s'accolent à la fin des formes verbales avec lesquelles ils fonctionnent : (*referirse, referirnos* versus *se référer, nous référer*). Ce mécanisme qui est à l'origine de la formation d'un grand nombre de formes graphiques différentes, complique par ailleurs le repérage direct des pronoms personnels en tant que tels.[1]

Les langues à déclinaisons dispersent les différents cas d'un même substantif en un plus ou moins grand nombre de graphies différentes. Ainsi, le russe éparpille-t-il en plusieurs formes graphiques les flexions d'un même substantif selon la fonction grammaticale qu'il remplit dans la phrase, faisant par ailleurs l'économie d'un grand nombre de prépositions.

ministry	les ministres
sovet ministrov	le conseil des ministres
predsedatelh soveta ministrov	le président du conseil des ministres
predsedatelå soveta ministrov	du président du conseil des ministres

Selon l'intérêt que l'on attend du repérage des catégories grammaticales d'un texte, pour l'analyse lexicométrique, on considérera cette circonstance comme particulièrement favorable ou, au contraire, comme une gêne considérable pour le repérage des différentes flexions d'une même unité textuelle.

[1] Cf. sur ce sujet particulier Romeu (1992).

En sus des déclinaisons, l'allemand présente, par rapport aux langues que nous venons de mentionner, la particularité, de créer des mots composés en agglomérant plusieurs substantifs, ou un radical verbal et un substantif.

Trinkwasser	eau potable,
Zusammengehörigkeitsgefühl	sentiment d'appartenance à

L'ordre d'agglomération des substantifs en allemand étant l'inverse de l'ordre des mots dans les composés français, un tri par la fin permet de repérer de nombreux composés ayant le même radical de base.
Lorsque l'on considère, par exemple, la liste des noms composés que l'on peut prélever dans un texte donné et qui contiennent la forme (*Bildung=formation*), la nécessité de scinder chacune de ces unités composée en fragments plus élémentaires avant de soumettre les textes à l'analyse lexicométrique paraît beaucoup moins évidente.

Bildung	formation, forme, culture
Hochschulbildung	enseignement des grandes écoles
Weiterbildung	formation continue
Preisbildung	formation des prix
Ausbildung	culture
Wortbildung	morphologie (lexicale)
Gesichtsbildung	forme du visage
Herzensbildung	noblesse d'âme

D'autant que l'on trouvera dans ces mêmes textes des mots composés pour lesquels la forme se trouve placée au début du mot composé.

Les expériences que nous avons pu réaliser à partir de textes rédigés dans différentes langues ont montré que, la plupart du temps, les particularités morpho-syntaxiques de chacune des langues concernées ne constituaient pas un obstacle majeur à l'approche des textes par les méthodes de la statistique textuelle. Comme on le verra plus loin, les typologies réalisées à partir des décomptes textuels se révèlent peu sensibles aux variations de l'unité de décompte.

Répétons-le, la forme graphique ne constitue en aucun cas une unité *naturelle* pour le dépouillement des textes ; l'avantage des décomptes en formes graphiques réside avant tout dans la facilité incomparable qu'il y a à les automatiser.

2.2 Segmentation et numérisation d'un texte

La mise en oeuvre des traitements informatisés peut être grandement simplifiée par l'application d'une technique de base appelée : la *numérisation du texte.*

Cette technique consiste à faire abstraction, pendant l'étape des calculs, de l'orthographe des formes décomptées pour ne retenir qu'un numéro d'ordre qui sera associé à toutes les occurrences de cette forme. Ces numéros seront stockés dans un dictionnaire de formes, propre à chaque exploitation. Ce dernier permettra, à l'issue des calculs, de reconstituer le graphisme des formes du texte mises en évidence par les calculs statistiques.

Tableau 2.1

Exemples de réponses à la question *Enfants*

Ind=01	*les difficultés financières et matérielles*
Ind=02	*les problèmes matériels, une certaine angoisse vis à vis de l'avenir*
Ind=03	*la peur du futur, la souffrance, la mort, le manque d'argent*
Ind=04	*l'avenir incertain, les problèmes financiers*
Ind=05	*les difficultés financières*
Ind=06	*les raisons matérielles et l'avenir qui les attend*
Ind=07	*des problèmes financiers*
Ind=08	*l'avenir difficile qui se prepare, la peur du chomage*
Ind=09	*l'insecurite de l'avenir*
Ind=10	*le manque d'argent*
Ind=11	*la guerre éventuelle*
Ind=12	*la charge financière que ca représente, la responsabilité morale aussi*
Ind=13	*la situation économique, quand le couple ou la femme n'est pas psychologiquement prêt pour accueillir un enfant*
Ind=14	*raisons éthiques*
Ind=15	*les problèmes de chômage, les problèmes d'ordre matériel*
Ind=16	*le fait de l'insécurité face au futur, la peur des responsabilités que cela implique*
Ind=17	*l'égoïsme (un chien vaut toujours mieux qu'un enfant), le manque d'argent faussement interprété*
Ind=18	*à Paris les conditions de logement, c'est fatigant, ca amène souvent à interrompre son travail, dans la situation mondiale actuelle c'est problématique*
Ind=19	*l'incertitude de l'avenir, les difficultés sociales*
Ind=20	*les raisons pécuniaires, la crainte de l'avenir à tous les niveaux*
Ind=21	*le chômage, les menaces de guerre*
Ind=22	*le refus de l'énorme responsabilité qui consiste à mettre un enfant au monde*
Ind=23	*les difficultés financières, la peur de s'investir affectivement*
Ind=24	*les raisons financières, l'augmentation du chômage*
Ind=25	*les difficultés financières*

Ind=26 *les problèmes d'argent surtout la peur du lendemain*
Ind=27 *l'argent, le manque d'argent*
Ind=28 *la trop grande responsabilité morale*
Ind=29 *le manque d'argent des raisons de santé parfois*
Ind=30 *vouloir garder une certaine indépendance, faibles ressources, conditions de logement*

La question ouverte a pour libellé : *Quelles sont les raisons qui, selon vous, peuvent faire hésiter une femme ou un couple à avoir un enfant ?*

Elle a été posée en 1981 à 2000 personnes représentant la population des résidents métropolitains de 18 ans et plus. Pour plus d'informations sur l'enquête utilisée (enquête sur les conditions de vie et aspirations des Français), cf. Lebart et Houzel van Effenterre (1980), Babeau et al. (1984), Lebart (1982b, 1987).

La numérisation du texte s'appuie sur des critères formels tels que : l'ordre alphabétique, le nombre des caractères qui composent la forme, etc. Dans ce qui suit nous avons choisi de numéroter les formes du texte en utilisant une combinaison de ces critères.

Les formes sont d'abord triées en fonction de leur fréquence dans l'ensemble du corpus. L'ordre alphabétique départage les formes de même fréquence.

On donnera un exemple de numérisation d'un corpus de réponses à une question ouverte (le corpus ***Enfants***), puis on introduira un corpus **P**, (comme *pédagogique*) de dimensions très réduites, qui servira par la suite à introduire plusieurs notions et définitions.

2.2.1 Numérisation sur le corpus *Enfants*

Dans le corpus constitué par les 2000 réponses à la question ouverte ***Enfants*** dont un extrait figure sur le tableau 2.1, la forme *de* est la plus fréquente avec 915 occurrences.

La forme *les* vient en 5ème position avec 442 occurrences. Les formes*: problèmes* (108 occ.)*, l'* (674 occ.) et *avenir* (318 occ.) ont respectivement les numéros : 31, 2 et 8. La numérisation de la séquence :

les problèmes de l'avenir, les problèmes financiers.

que nous avons vue plus haut donnera donc la séquence numérique :

5 31 1 2 8 , 5 31 39

Tableau 2.2

Numérisation de 10 réponses à la question *Enfants* présentées au tableau 2.1

```
1  *   5  42  19  12  56
2  *   5  31 230  28 198   283 138     1  2   8
3  *   3  23  24 166   3  1251   3  1074  4  22  6
13
4  *   2   8  70   5  31    39
5  *   5  42  19
6  *   5  42  19   5  17    56  12     2  8  47  5
716
7  *  11  31  39
8  *   2   8  88  47  81  1147   3    23 24   9
9  *   2  52   1   2   8
10 *   4  22   6  13
```

On trouve au tableau 2.2 la numérisation effectuée à partir des premières réponses présentées au tableau 2.1.

Ainsi, les traitements lexicométriques réalisés à partir des séquences textuelles numérisées vont se trouver simplifiés. Par ailleurs, le volume du stockage du texte en mémoire d'ordinateur se trouve considérablement réduit.

Notons ici que certains logiciels considèrent les ponctuations comme des formes à part entière, d'autres les ignorent complètement. On leur conférera ici un statut spécial de caractères délimiteurs.

2.2.2 Le corpus P

Pour de simples illustrations des définitions données dans ce chapitre, nous travaillerons sur un "texte" appelé, dans la suite de ce chapitre le "corpus" **P**, où les lettres capitales représentent des formes dont ni le graphisme exact, ni la signification n'ont d'importance pour ce qui suit. Il s'agit d'un "texte" simple, qui fournit cependant quelques exemples de redondances que nous utiliserons plus loin.

Tableau 2.3

Le "corpus" pédagogique P

A	B	C	A	C ;	C	D	E	A	B	C .	D	F.	
(1	2	3	4	5)	(6	7	8	9	10	11)	(12	13)	

Ce "corpus" compte treize *occurrences,* numérotées de 1 à 13 ; sa taille, que l'on désignera par la lettre *T*, vaut :

$$T = 13.$$

Parmi ces occurrences, il en est qui sont des occurrences de la même forme graphique. Ainsi, les occurrences qui portent les numéros : 1, 4 et 9 sont toutes des occurrences de la forme A.

Au total on dénombre dans ce corpus six formes différentes : A, B, C, D, E, F. Ces formes constituent le vocabulaire du corpus P. Désignant par *V* l'effectif du vocabulaire, nous écrirons dans le cas présent :

$$V = 6.$$

2.3 L'étude quantitative du vocabulaire

La lexicométrie comprend des méthodes qui permettent d'opérer des réorganisations formelles de la séquence textuelle et aussi de procéder à des analyses statistiques portant sur le vocabulaire à partir d'une segmentation.

En amont de ces phases de travail, il est possible de calculer à partir du texte numérisé un certain nombre de paramètres distributionnels qui relèvent en quelque sorte d'une application et d'une adaptation de la statistique descriptive élémentaire aux particularités des recueils textuels.

2.3.1 Fréquences, gamme des fréquences

Revenons à l'exemple du corpus **P**. La forme A possède trois occurrences : elle a une *fréquence* égale à trois, les *adresses* de cette forme dans le corpus sont : 1, 4 et 9. L'ensemble des adresses d'une forme constitue sa *localisation*.

Dans ce corpus la forme C est, avec ses 4 occurrences, la forme la plus fréquente ; la *fréquence maximale* est donc égale à 4. Ce que l'on note habituellement :

$$F_{max} = 4$$

La forme E n'apparaît qu'une seule fois dans le texte ; c'est une forme de fréquence 1 ou encore un *hapax*[1]. Dans ce corpus les hapax sont au nombre de deux : les formes E et F. Désignant par V_1 l'effectif des formes de fréquence 1 on écrira :

$$V_1 = 2$$

On notera de la même manière : V_2, V_3, V_4, … les effectifs qui correspondent respectivement aux fréquences deux, trois et quatre. Dans notre corpus :

$$V_2 = 2 \; ; \quad V_3 = 1 \; ; \quad V_4 = 1 \; .$$

On peut également représenter la *gamme des fréquences* de la manière suivante :

Fréquence	:	*1*	*2*	*3*	. . .	*Fmax*
Effectif	:	V_1	V_2	V_3	. . .	V_{Fmax}

Pour le corpus P, la gamme des fréquences s'écrit :

Fréquence	:	*1*	*2*	*3*	*4*
Effectif	:	*2*	*2*	*1*	*1*

Les fréquences et les effectifs correspondant à chacune de ces fréquences sont liés entre eux par des relations d'ensemble. Certaines de ces relations sont simples. La somme des effectifs correspondant à chacune des fréquences est égale au nombre des formes contenues dans le corpus, ce que l'on note :

$$\sum_{i=1}^{Fmax} V_i = V$$

D'autre part, la somme des produits (fréquence x effectif) pour toutes les fréquences comprises entre 1 et *Fmax*, bornes incluses, est égale à la longueur du corpus.

$$\sum_{i=1}^{Fmax} V_i \times i = T$$

[1] Du grec *hapax legomenon* , "chose dite une seule fois".

2.3.2 La loi de Zipf

Si le corpus que l'on dépouille compte quelques milliers d'occurrences, on observe que la gamme des fréquences acquiert des caractéristiques plus délicates à décrire. Ces caractéristiques sont associées au nom de G. K. Zipf.

La première découverte[1] de Zipf fut de constater que toutes les gammes de fréquences obtenues à partir de corpus de textes présentent des caractéristiques communes.

Très grossièrement, on peut dire que si l'on numérote les éléments de la gamme des fréquences préalablement rangés dans l'ordre décroissant, on aperçoit une relation approximative entre le rang et la fréquence. On dit parfois :

"le produit (rang x fréquence) est à peu près constant"

Ainsi dans l'*Ulysses* de Joyce qui compte 260000 occurrences les décomptes opérés dès 1937 par H.L. Handley (cf. Guilbaud, 1980) donnent par exemple :

au rang	10	la fréquence	2653
au rang	100	la fréquence	265
au rang	1000	la fréquence	26
au rang	10000	la fréquence	2

Le diagramme de Pareto fournit une représentation très synthétique des renseignements contenus dans la gamme des fréquences. Ce diagramme est constitué par un ensemble de points tracés dans un repère cartésien.

Sur l'axe vertical, gradué selon une échelle logarithmique, on porte la fréquence de répétition *F*, qui varie donc de 1 à *Fmax*, la fréquence maximale du corpus. Sur l'axe horizontal, gradué selon la même échelle logarithmique, on porte, pour chacune des valeurs de la fréquence *F* comprises entre 1 et *Fmax*, le nombre *N(F)* des formes répétées au moins *F* fois dans le corpus. La courbe obtenue est donc une courbe cumulée.

De nombreuses expériences faites dans le domaine lexicométrique montrent que, quel que soit le corpus de textes considéré, quelle que soit la norme de dépouillement retenue, les points ainsi tracés s'alignent approximativement le long d'une ligne droite. Le diagramme de Pareto est l'une des formes sous

[1] Mandelbrot (1961), dans un exposé reprenant les diverses tentatives de modélisation de la gamme des fréquences, souligne les apports de J.B. Estoup (op. cit.) dans ce domaine.

laquelle peut s'exprimer une loi très générale touchant à l'économie du vocabulaire dans un corpus de textes.

L'étude de la forme générale de la courbe ainsi obtenue, des "meilleurs" ajustements possibles à l'aide de courbes définies par un petit nombre de paramètres, a longtemps retenu l'attention des chercheurs. On trouve la trace de préoccupations similaires dans la plupart des ouvrages qui traitent de statistique lexicale[1]. L'enjeu principal de ces débats est la détermination d'un modèle théorique très général de la gamme des fréquences d'un corpus. L'élaboration d'un tel modèle permettrait en effet d'interpréter par la suite, en termes de stylistique, les écarts particuliers constatés empiriquement sur chaque corpus.

Malheureusement la démonstration faite par chaque auteur de l'adéquation du modèle qu'il propose s'appuie, la plupart du temps, sur des dépouillements réalisés selon des normes qui varient d'une étude à l'autre. Par ailleurs, la présentation de ces "lois théoriques", dont la formule analytique est souvent complexe, apporte peu de renseignements sur les raisons profondes qui sont à l'origine de ce phénomène.

La figure 2.2 donnera un exemple de diagrammes de Pareto permettant de comparer des textes traditionnels avec des corpus formés de juxtaposition de réponses à des questions ouvertes.

2.3.3 Mesures de la richesse du vocabulaire.

Considérons maintenant le "texte" constitué par la réunion des 2 000 réponses à la question ***Enfants*** évoquée au début de ce chapitre (cf. tableau 2.1). Dans ce qui suit on désignera cet ensemble de réponses par le symbole Q_1. L'ensemble Q_1 compte 15 523 occurrences. Le nombre des formes *V* est égal à 1 305.

On appellera Q_2 un ensemble du même type, constitué par la réunion des réponses à une autre question ouverte de ce même questionnaire :

Etes-vous d'accord avec l'idée suivante : La famille est le seul endroit où l'on se sente bien et détendu ? Pouvez-vous dire pourquoi ?

Ce second ensemble compte 22 202 occurrences pour 1651 formes. Ces données numériques n'ont de valeur que si on les compare à d'autres données du même type. On peut alors caractériser chacune des questions de l'enquête par le nombre total des occurrences présentes dans les réponses des enquêtés

[1] Cf., par exemple, Mandelbrot (1968), Petruszewycz (1973), Muller (1977), Ménard (1983), Labbé et al. (1988).

ou encore par le nombre des formes différentes dans chacun de ces ensembles.

On n'entreprendra pas ici l'analyse comparée de ces comptages pour les différentes questions posées lors de l'enquête précitée. On se bornera à comparer les principales caractéristiques lexicométriques pour Q_1 et Q_2 avec d'autres textes prélevés de manière moins artificielle dans des textes rédigés par un même locuteur.

Dans le tableau 2.4 on trouve ces caractéristiques lexicométriques calculées pour six "textes" provenant d'horizons très divers.

Les deux premiers échantillons correspondent aux questions Q_1 et Q_2 que nous venons de décrire.

Les deux suivants, que l'on désignera par S_1 et S_2, sont des fragments de 20 000 occurrences prélevés dans un corpus de textes socio-politiques[1].

Les deux derniers G1 et G2 sont des prélèvements de 20 000 occurrences consécutives découpés dans un roman français contemporain[2].

Tableau 2.4

Les principales caractéristiques lexicométriques pour deux question d'un questionnaire socio-économique (Q1, Q2), deux fragments de textes syndicaux (S1, S2) et deux fragments de texte littéraire (G1, G2).

	occ.	formes	hapax	fmax	
Q1	15 523	1 305	648	*de*	915
Q2	22 202	1 651	881	*on*	1 128
S1	20 000	3 061	1 576	*de*	1 180
S2	20 000	3 233	1 642	*de*	1 238
G1	20 000	4 605	3 003	*de*	968
G2	20 000	4 397	2 831	*de*	908

Ces quelques données ne prétendent nullement résumer à elles seules l'éventail des situations possibles. Elles aideront seulement à situer très

[1] Ce corpus réuni par J. Lefèvre et M. Tournier au laboratoire de St.Cloud, comprend les résolutions adoptées dans leurs congrès respectifs par les principales centrales syndicales ouvrières françaises au cours de la période 1971-1976. L'étude de ce corpus de textes syndicaux a donné lieu à de nombreuses publications (cf., notamment, Bergounioux et al., 1982.)

[2] Il s'agit du roman de Julien Gracq : *Le Rivage des Syrtes*, Paris , Corti, 1952.

grossièrement les caractéristiques lexicométriques des textes constitués par la concaténation de réponses à une question ouverte parmi d'autres exemples de comptages pratiqués sur des textes.

Comme on le voit, les "textes" Q1 et Q2 sont beaucoup moins riches en formes que les échantillons de textes syndicaux eux-mêmes beaucoup moins riches que les fragments de texte littéraire.

Ceci est lié, entre autres, au fait que les textes Q1 et Q2 comptent beaucoup moins d'hapax que les quatre autres fragments présentés au tableau 2.5.

Pour caractériser plus précisément la gamme des fréquences du texte Q1 par rapport au texte S1, on utilisera le diagramme de Pareto présenté plus haut.

On voit sur la figure 2.2 (et sur le tableau des fréquences maximales) que les fréquences maximales (i.e. fréquence de la forme la plus fréquente) sont proches pour les deux textes bien que Q1 compte 5 000 occurrences de moins que S1.

Les formes de faible fréquence sont plus nombreuses dans le texte S1. Cet avantage s'estompe à mesure que l'on se rapproche de la fréquence 20.

Entre la fréquence 20 et la fréquence 100, la courbe relative au texte Q1 est au-dessus de la courbe qui correspond au texte S1. Ceci traduit un excédent de formes de fréquence moyenne pour la question Q1. Entre la fréquence 100 et la fréquence maximale, la courbe correspondant au texte S1 reprend l'avantage, ce qui traduit un excédent relatif des formes dont la fréquence est comprise dans cet intervalle par rapport à la question Q1.

On retiendra donc que le texte Q1 possède un vocabulaire plus réduit et une répétitivité plus grande dans l'intervalle des fréquences moyennes que son homologue S1.

Guide de lecture du Diagramme de Pareto

Sur l'axe vertical, gradué selon une échelle logarithmique, on lit la fréquence de répétition *F*.

Sur l'axe horizontal, gradué de la même manière, on trouve, pour chaque fréquence *F* comprise entre *1* et *Fmax*, le nombre des formes répétées au moins *F* fois dans le corpus.

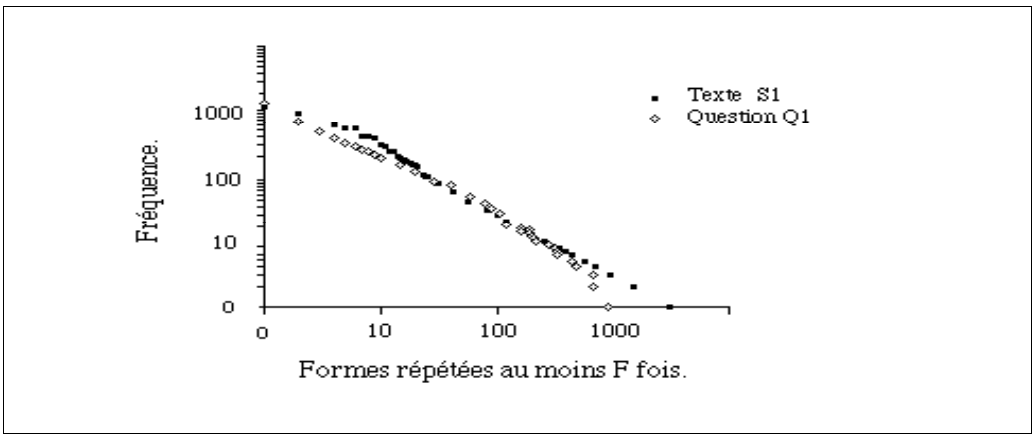

Figure 2.2

Diagramme de Pareto pour les "textes" S_1 (20 000 occ. de texte suivi) et Q_1 (15 000 occ. prises dans des réponses à une question ouverte).

2.4 Documents lexicométriques

Plusieurs documents permettent, à partir d'une simple réorganisation de la séquence textuelle, de porter sur le corpus des textes que l'on étudie un regard différent de celui qui résulte de la simple lecture séquentielle. Il faut noter que l'habitude de produire de tels documents est beaucoup plus ancienne que l'analyse lexicométrique assistée par ordinateur. Cependant pour automatiser entièrement la confection de tels états il est indispensable de formaliser l'ensemble des règles qui président à leur établissement.

2.4.1 Index d'un corpus

Les index, qui constituent par rapport au corpus de départ une réorganisation des formes et des occurrences, permettent de repérer immédiatement, pour chacune des formes, tous les endroits du corpus où sont situées ses occurrences.

Pour retrouver ces occurrences dans le texte de départ, on utilise généralement, de préférence au système des adresses, un système de coordonnées numériques plus étroitement liées à l'édition du texte comme : le tome, la page, la ligne, la position de l'occurrence dans la ligne, etc. Ces renseignements qui permettent de retourner plus facilement au document d'origine sont les *références* associées à chacune des occurrences.

Dans les index, les formes peuvent être classées selon des critères différents. On appelle *index alphabétique* un index dans lequel les formes sont classées

selon l'ordre lexicographique (i.e. l'ordre alphabétique couramment utilisé dans les dictionnaires).

On appelle *index hiérarchique* un index dans lequel les formes sont rangées en premier lieu par ordre de fréquence décroissante. Dans ce type d'index, on départage en général les formes de même fréquence d'après l'ordre lexicographique. L'ordre que l'on vient de définir est l'ordre lexicométrique.

Tableau 2.5

L'index hiérarchique du corpus P

fréquence	*forme*	*références*			
4	C	3	5	6	11
3	A	1	4	9	
2	B	2	10		
2	D	7	12		
1	E	8			
1	F	13			

Le tableau 2.5 contient l'index hiérarchique réalisé à partir du corpus **P**. Dans cet index, c'est le numéro d'ordre dans le corpus qui sert de référence pour chacune des occurrences.

2.4.2 Contextes, concordances

Pour chacune des formes du corpus, on peut, en se servant des index, localiser l'ensemble de ses occurrences dans le texte d'origine. Ce travail effectué, il peut être intéressant d'étudier systématiquement les contextes immédiats dont ces occurrences ont été extraites. Cependant pour une forme pourvue d'une fréquence élevée, cette opération, qui exige un va-et-vient constant entre le texte et l'index, se révèle vite fastidieuse.

Il est possible d'opérer des réorganisations des formes et des occurrences du texte : les occurrences d'une même forme se trouveront rassemblées en un même endroit accompagnées d'un petit fragment de contexte immédiat dont on fixera la longueur en fonction des besoins particuliers.

Ici encore, la manière de procéder a largement précédé l'apparition de l'ordinateur dans le domaine des études textuelles. Le recours à la machine permet simplement d'obtenir en quelques secondes et avec des risques

d'erreur considérablement réduits des états dont la confection suffisait parfois à occuper nos ancêtres leur vie durant.[1]

Tableau 2.6

Concordance de la forme C dans le corpus P.

contexte avant	*forme-pôle*	*contexte après*	*référence*
A B	**C**	A C ;C D E A B	3
A B C A	**C**	; C D E A B C .D	5
A B C A C;	**C**	D E A B C .D F	6
C D E A B	**C**	. D F.	11

(Rappel du corpus P) **:**

A B **C** A **C** ; **C** D E A B **C** . D F.

On appellera *forme-pôle* la forme dont on regroupera les contextes. On trouve ci-dessus l'exemple d'une concordance qui regroupe toutes les occurrences relatives à une même forme-pôle, la forme C du corpus P.

Dans l'exemple qui suit, les lignes de la concordance sont triées d'après l'ordre alphabétique de la forme qui suit la forme-pôle (successivement A, C, D, D). Comme plus haut, la référence située à droite est le numéro d'occurrence dans le texte.
Malgré leur volume, les concordances se révèlent à l'usage des instruments très pratiques pour l'étude des textes. En effet, grâce aux réorganisations de la séquence textuelle qu'elles permettent, les concordances fournissent, sur l'emploi d'une forme donnée, une vision plus synthétique que celle qui résulte de la lecture séquentielle. En particulier, elles permettent d'étudier plus facilement les rapports qui peuvent exister entre les différents contextes d'une même forme.

On trouve au tableau 2.7 les premières lignes d'une *concordance* réalisée pour les contextes de la forme : *enfants* dans le "texte" Q1. La forme *enfants* compte 148 occurrences parmi l'ensemble des réponses. C'est dire que la concordance complète pour cette forme compte 148 lignes. Les lignes de contexte sont ici triées par ordre alphabétique croissant en fonction de la forme qui suit la forme-pôle.

[1] On trouvera des éléments sur l'histoire des concordances dans Sekhraoui (1981).

Remarquons que cette réorganisation des occurrences du texte initial fait apparaître des liens *(mettre les/des enfants au monde, enfants en bas âge)* qui auraient peut-être échappé à une simple lecture séquentielle du texte.

Tableau 2.7
Concordance de la forme *enfants* dans le corpus des réponses à la question *Enfant*

```
L 304   c' est pas drole d' avoir des enfants a l' heure actuelle a
L 420   la peur de l' avenir pour les enfants a quelle place peuvent
L2415  onomique* risques d' avoir des enfants anormaux* avenir incer
L2281  s raisons medicales( risque d' enfants anormaux)* manque de c
L1760  est pas la peine de mettre des enfants au monde dans un clima
L2307  c' est bien beau de mettre des enfants au monde mais si c' es
L1879  chomage, on ne peut mettre les enfants au monde sans leur ass
L 429  affection entre le mari et les enfants* aucune garantie d' av
L 710  as* les gens n' aiment pas les enfants* aucune raison n' est
L1461  l' age, l' argent, la peur des enfants aussi, le confort a de
L1295  ou d' argent* ne pas aimer les enfants, avoir peur d' en avoi
L2165  ur, elles sont tributaires des enfants c' est astreignant* le
L 498  and on n' en a pas assez, deux enfants c' est le maximum* les
L1009   il y en a qui n' aime pas les enfants, c' est le travail, la
L 846   desirent pas, un mariage sans enfants, c' est pas un mariage
L2013  ve au chomage avec beaucoup d' enfants c' est plus difficile*
L1903   la vie et le chomage pour les enfants* c' est trop dur d' el
L1647  er une meilleure education aux enfants, ca coute cher actuell
L 285  st difficile de s' occuper des enfants* ca depend de la menta
L1720  roblemes pour faire garder les enfants, ca permet au couple d
L2170  ieres, on choisit d' avoir des enfants, choix delibere de la
L  68   financieres, ne pas aimer les enfants* chomage* difficultes
L1839  probleme financier, avenir des enfants (chomage)* le cout de
L1196  rs, peur de l' avenir pour les enfants* crise actuelle peur d
L 921  fants, difficile d' elever des enfants de nos jours car les e
L 441   l' avenir du chomage pour les enfants* des femmes ne sont pa
L 921   REP* peur du chomage pour ses enfants, difficile d' elever d
L 289  les, materielles, la garde des enfants difficile due au trava
L2072  P* la peur du chomage pour ses enfants* difficulte d' assumer
L 966  eulent plus se priver pour les enfants* economique, pas de de
L 744   couples qui n' aiment pas les enfants* egoisme* de peur d' a
L 612  mage* la femme ne veut plus d' enfants, elle veut sa liberte,
L1650  structures d' accueil pour les enfants en bas age dans le cas
L1467   actuelle n' est fait pour les enfants en bas age les infrast
L 166  nconvenients presentes par des enfants en bas age( trop absor
```

Remarquons également qu'un tri réalisé en fonction de l'occurrence qui précède la forme-pôle aurait eu pour effet de disperser les occurrences de ces expressions à différents endroits de la concordance et d'en rassembler d'autres qui se trouvent dispersées par le mode de tri adopté ici.

2.4.3 L'accroissement du vocabulaire

Lorsqu'on ajoute des occurrences à un texte, son vocabulaire (le nombre de formes distinctes) a tendance à augmenter. L'étude de l'accroissement du vocabulaire au fil des occurrences d'un même texte constitue l'un des domaines traditionnels de la statistique lexicale.

Pour le corpus P, cet accroissement est décrit de la manière suivante :

	A	B	C	A	C	;	C	D	E	A	B	C	.	D	F.
occurrences	1	2	3	4	5		6	7	8	9	10	11		12	13
formes	1	2	3	3	3		3	4	5	5	5	5		5	6

Pour des corpus de plus grande taille, l'accroissement du vocabulaire subit une double influence.

a) La prise en compte de nouvelles occurrences tend à augmenter le nombre total des formes du corpus (plus il y a d'occurrences, plus les formes distinctes sont nombreuses).

b) Lorsque la taille du corpus augmente, le taux des formes nouvelles apportées par chaque accroissement du nombre des occurrences a tendance à diminuer.

Il s'ensuit que le nombre des formes contenues dans un corpus n'est pas proportionnel à sa taille.

On appellera : *grandeurs lexicométriques de type T* les grandeurs qui croissent *à peu près* en proportion de la longueur du texte. Comme on le verra plus loin les fréquences des formes les plus fréquentes d'un corpus sont des grandeurs de ce type.

On appellera *grandeurs lexicométriques de type V* les variables, telles le nombre des formes, dont l'accroissement a tendance à diminuer avec la longueur du texte.

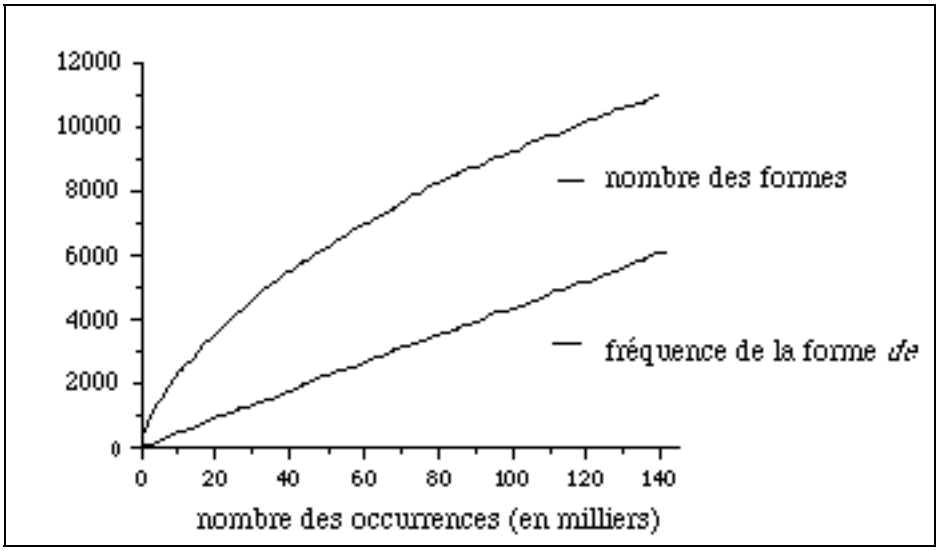

Figure 2.3

Le nombre des formes et le nombre des occurrences de la forme *de* en fonction de la longueur du corpus *Duchesne*.

La figure 2.3 fournit un exemple de tels comptages réalisés sur une grande échelle[1]. Le corpus sur lequel ont été effectués ces comptages compte 11 070 formes pour 141 182 occurrences.

Comme on le voit sur cette figure, le nombre des occurrences de la forme *de*, forme la plus fréquente du corpus, croît linéairement alors que l'accroissement du nombre des formes tend à diminuer au fur et à mesure que le corpus s'allonge.

2.4.4 Partitions du corpus

Le nombre des occurrences dans une partie du corpus est la longueur de la partie. On notera t_j la longueur de la partie numéro j.

Si chaque occurrence du corpus est affectée à une partie et une seule, la somme des longueurs de parties est égale à la longueur du corpus.

$$\sum_{i=1}^{n} t_i = T$$

[1] Ce corpus réuni par J.Guilhaumou (1986) comprend 96 livraisons d'un journal paru pendant la révolution française : *Le Père Duchesne* de Jacques Hébert.

Dans un corpus divisé en *N* parties, correspondant par exemple à *N* périodes ou à *N* locuteurs différents, on parlera des sous-fréquences d'une forme dans chacune des parties.

La suite de *N* nombres constituée par la succession des sous-fréquences prises dans l'ordre des parties est la ventilation des occurrences de cette forme, ou plus simplement la *ventilation* de cette forme, dans les parties du corpus.

La répartition d'une forme est égale au nombre des parties du corpus dans lesquelles elle est présente. Si une forme n'est attestée que dans une seule des parties du corpus elle appartient au *vocabulaire original* de cette partie. Si une forme appartient au vocabulaire de chacune des parties du corpus c'est une forme *commune*.

On notera que toutes ces qualités dépendent très étroitement de la partition opérée sur le corpus de départ. Sur un même corpus, en effet, on sera souvent conduit, en fonction d'objectifs de recherche particuliers, à opérer plusieurs partitions qui serviront de base à des analyses différentes.

Si l'on divise, par exemple, le corpus **P** en trois parties : P1, P2, P3 délimitées par les deux signes de ponctuation situés respectivement après les occurrences 5 et 11, on peut calculer les trois sous-fréquences de la forme A et constituer sa *ventilation* dans les parties du corpus.

Cette ventilation s'écrit :

A : (2 1 0)

La forme A n'est attestée que dans deux des trois parties du corpus **P**. La forme F appartient au *vocabulaire original* de la partie P3.

2.4.5 Tableaux lexicaux

Il est commode de ranger les décomptes des occurrences de chacune des *V* formes dans chacune des *N* parties du corpus dans un tableau rectangulaire qui compte *V* lignes et *N* colonnes. Dans ce tableau, les formes sont classées de haut en bas selon l'ordre lexicométrique. On appelle ce tableau : le *Tableau Lexical Entier* (TLE). Les nombres *V* et *N* sont les dimensions du TLE. Des tableaux lexicaux partiels peuvent être construits, par exemple en ne retenant que les formes dont la fréquence est au moins égale à un certain seuil.

La partition en trois parties du corpus P permet de construire un TLE de dimensions (5 x 3). La ligne *longueur des parties* qui figure sous le TLE est la marge horizontale du tableau. On obtient chacun des nombres situés sur cette

ligne en additionnant tous les nombres qui se trouvent dans la colonne située au-dessus de lui. La marge horizontale du tableau indique donc bien la longueur de chacune des parties.

Tableau 2.8

Le Tableau Lexical Entier (TLE) du corpus P muni de la partition en 3 classes et ses marges.

forme	P1	P2	P3	*fréquence des formes*
C	2	2	0	4
A	2	1	0	3
B	1	1	0	2
D	0	1	1	2
E	0	1	0	1
F	0	0	1	1
longueur des parties	5	6	2	13

Sur la droite du TLE la colonne *fréquence des formes* est la marge verticale. On obtient chacun des nombres situés sur cette colonne en additionnant les nombres du TLE situés sur la même ligne que lui.

Les nombres qui forment la marge verticale du tableau, indiquent donc la fréquence de chacune des formes dans le corpus. On trouvera au chapitre 3 (tableau 3.1) et surtout au chapitre 5 (tableau 5.3) des exemples de tableaux lexicaux construits à partir des données d'enquêtes.

2.5 Les segments répétés

Les formes graphiques ne sont évidemment pas des unités de sens : même si l'on écarte celles qui ont des fonctions purement grammaticales, elles apparaissent dans des mots composés, dans des locutions ou expressions qui infléchissent, voire modifient totalement leurs significations. En effet, on observe les récurrences d'unités comme : *sécurité sociale, niveau de vie, etc.* dotées en général d'un sens qui leur est propre et que l'on ne peut déduire à partir du sens des formes qui entrent dans leur composition.

Dans les études textuelles, il sera utile de compléter les résultats obtenus à partir des décomptes de formes graphiques par des comptages portant sur des unités plus larges, composées de plusieurs formes.
Il est alors utile de soumettre ces unités aux mêmes traitements statistiques que les formes graphiques. Mais ces décisions de regroupement de certaines occurrences de formes graphiques ne peuvent être prises a priori sans entacher de subjectivité les règles du dépouillement.

Dans ce paragraphe on introduira des objets textuels de type nouveau : *les segments répétés*[1]. On montrera ensuite au cours des chapitres suivants comment utiliser les produits de cette formalisation dans les différentes méthodes d'analyses statistiques.

2.5.1 Phrases, séquences

La distinction des caractères du texte en caractères délimiteurs et non-délimiteurs (paragraphe 2.1.2) permet de définir une série de descripteurs relatifs aux formes simples. Pour aborder la description des segments, composés de plusieurs formes et répétés dans un corpus de textes, il s'avère utile de préciser le statut de chacun des caractères délimiteurs.

Si l'on se contente, en effet, de recenser toutes les suites de formes identiques, sans tenir compte des signes de ponctuation qui peuvent venir s'intercaler entre ces dernières, on sera amené à considérer comme deux occurrences d'un même segment deux suites de formes telles par exemple : [A, B, C D] et [A B C D] ce qui peut devenir une source de confusion.

Pour éviter d'avoir à pratiquer des décomptes séparés entre les occurrences de segments contenant des signes de ponctuation et celles des segments identiques aux signes de ponctuation près, on se bornera en général à la recherche des seuls segments ne chevauchant pas de ponctuation faible ou forte.

La question de la définition des limites de la phrase, réputée difficile chez les linguistes, ne trouve pas de solution originale dans le domaine lexicométrique. Pour les procédures formelles que l'on élabore, on se contentera, faute de mieux, de donner à certains des signes de ponctuation (le point, le point d'exclamation, le point d'interrogation) le statut de séparateur fort ou séparateur de phrase.

---

[1] Les procédures de repérage des segments répétés sont issues d'un travail réalisé à l'ENS de Fontenay-St.Cloud, cf. Lafon et Salem (1986), Salem (1984, 1986, 1987). Cf. aussi Bécue (1988).

Reprenons l'exemple du corpus **P** présenté plus haut :

```
A  B  C  A  C ; C  D  E  A  B  C . D  F.
1  2  3  4  5   6  7  8  9  10 11 12 13
```

On voit que le corpus P est composé de deux phrases.
Parmi les caractères délimiteurs nous choisirons également un sous-ensemble correspondant aux ponctuations faibles et fortes (en général : la virgule, le point-virgule, les deux points, le tiret, les guillemets et les parenthèses auxquelles on ajoutera les séparateurs de phrases introduits plus haut) que l'on appellera l'ensemble des délimiteurs de séquence. La suite des occurrences comprises entre deux délimiteurs de séquence est une séquence.

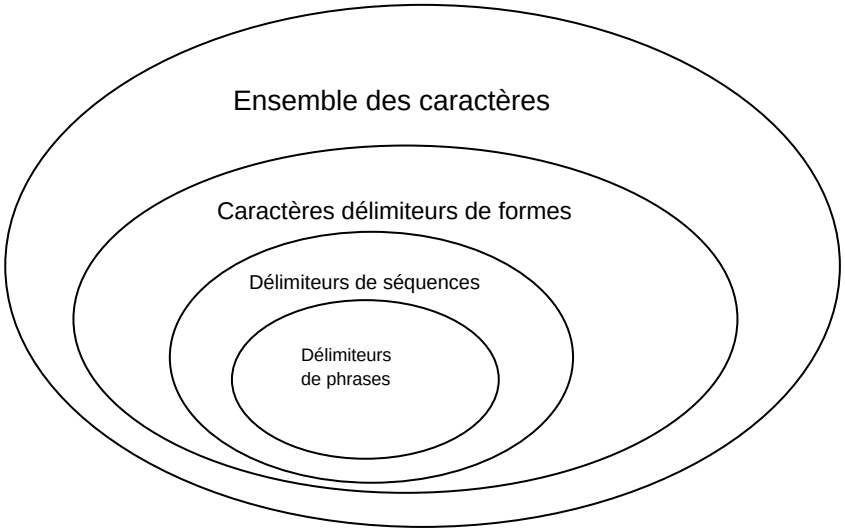

Figure 2.4

Classification des caractères.

Les délimiteurs de formes qui ne sont ni séparateurs de phrase ni délimiteurs de séquence seront, quant à eux, traités comme le caractère que l'on appelle blanc ou espace. On voit que le corpus P est composé de trois séquences. La classification des caractères opérée ci-dessus est résumée par la figure 2.4.

2.5.2 Segments, polyformes

Selon la définition donnée au paragraphe précédent, toutes les suites d'occurrences non séparées par un délimiteur de séquence sont des occurrences de segments, ou de polyformes. Les segments dont la fréquence est supérieure ou égale à 2 dans le corpus sont des segments répétés dans le corpus.

Reprenons le corpus P. Selon la définition que nous venons d'introduire, la première des séquences ABCAC du corpus P se décompose en segments de la manière suivante :

ABCAC, ABCA, BCAC, ABC, BCA, CAC, AB, BC, CA, AC.

Pour l'ensemble du corpus P, seuls les segments : ABC, AB, BC sont des segments répétés.

De la même manière, pour le "corpus" constitué par la réponse à la question Q_1 :

les problèmes financiers, les problèmes matériels

la première séquence contient les segments :

les problèmes financiers, *les problèmes*, *problèmes financiers*

Le segment de longueur 2 : *les problèmes* est le seul segment répété du corpus. On vérifie par ailleurs que toute séquence de longueur L contient obligatoirement: (L-1) segments de longueur 2, (L-2) segments de longueur 3, ... , etc.

Il est désormais difficile de distinguer les deux termes "segments" et "polyformes" qui fonctionnent presque comme des synonymes. On peut noter cependant qu'ils éclairent chacun de manière différente les objets que l'on étudie.

La dénomination de *segment* souligne bien le fait que ces unités sont obtenues en opérant des coupures dans un texte préexistant. Il faut se souvenir cependant que ces coupures sont pratiquées par un algorithme qui ne s'appuie que sur des critères formels tels que la longueur et l'absence de délimiteur de séquence intermédiaire[1].

Ainsi, dans la séquence : *le petit chat est mort*, les définitions retenues nous amènent à considérer, dans un premier temps du moins, le segment constitué par le syntagme nominal : *le petit chat* sur le même plan que le segment obtenu à partir des trois occurrences *petit chat est* qui ne constitue pas une unité *en langue*. Cependant, le repérage des segments les plus répétés dans un corpus de textes, en plus de l'éclairage quantitatif qu'il apporte sur ces problèmes, permet de mettre en évidence des unités qui constituent souvent des syntagmes autonomes.

A l'inverse, la dénomination de *polyforme* implique davantage l'idée qu'une suite de formes, dont on n'a pas encore nécessairement repéré l'existence ni a

[1] Les procédures de traitement lexicométrique ne peuvent, dans l'état actuel des choses, tenir compte des frontières de syntagmes (i.e. groupes de mots en séquence formant une unité à l'intérieur de la phrase) que le grammairien repère par une simple lecture.

fortiori toutes les occurrences éventuelles dans le texte que l'on étudie, possède, dans ce texte sinon en langue, une unité de fonctionnement qui lui est propre.

Ainsi, dans les textes socio-politiques, il est intéressant de localiser, en plus des occurrences des formes *sécurité* et *sociale*, les occurrences de la polyforme *sécurité sociale* qui fonctionne dans ces textes comme une unité qu'il est dommageable de scinder en deux formes isolées.

2.5.3 Quelques définitions et propriétés relatives aux segments

Comme dans le cas des formes, la fréquence d'un segment est le nombre des occurrences de ce segment dans l'ensemble du corpus de textes.

La terminologie employée pour décrire la localisation des formes simples s'applique sans difficulté à la localisation des segments, si l'on convient que l'adresse de chacune des occurrences d'un segment est donnée par celle de la première des occurrences de forme simple qui le composent. Ainsi, on dira sans risque de confusion que dans le corpus **P** le segment AB, de longueur 2 et de fréquence 2, possède deux occurrences dont les adresses sont respectivement : 1 et 9. La manipulation de la notion de segment exige cependant que soient définies et précisées quelques notions ou propriétés.

Eléments d'un segment, sous-segments

Le premier segment de longueur 3 du corpus **P**, le segment ABC, possède deux occurrences dans ce corpus. Les formes A, B et C sont, respectivement, les premier, deuxième et troisième éléments du segment ABC, les segments AB et BC les premier et deuxième sous-segments de longueur 2 de ce même segment.

Expansions, expansions récurrentes, expansions contraintes

Considérons le segment ABC. La forme C constitue pour le segment AB une *expansion droite* de longueur 1. On définit de la même manière l'expansion gauche d'un segment. Le segment ABC constitue pour le segment AB un *voisinage* de longueur 3.

Les deux occurrences du segment AB situées aux places 1 et 9 du texte P sont suivies de la même forme, en l'occurrence la forme C. On dira donc que dans ce corpus, la forme C est une *expansion récurrente* (droite) pour le segment AB. Le segment ABC est un *voisinage récurrent* pour le segment AB.

Dans la mesure où il n'existe aucune autre expansion du segment AB dans ce corpus, la présence d'une occurrence de AB implique obligatoirement la

présence d'une occurrence du segment ABC. On dira que C est une *expansion contrainte* (de longueur 1 à droite) pour le segment AB.

Segments contraints, segments libres

De la même manière les sous-segments d'un segment donné se divisent en deux catégories : les *segments contraints* et les *segments libres*. Un sous-segment est contraint (dans un autre segment S de longueur supérieure) si toutes ses occurrences correspondent à des occurrences du segment S.

Si au contraire un segment possède plusieurs expansions distinctes, qui ne sont pas forcément récurrentes, on dira que c'est un segment libre. Pour notre second exemple, *problèmes* possède une expansion récurrente à gauche, *les problèmes,* qui est de longueur deux. Ce segment est lui-même un segment contraint dans le corpus.

Tableau des segments répétés (TSR)

Comme dans le cas des formes simples, on rangera dans ces tableaux les comptages portant sur la ventilation des occurrences d'un même segment dans les différentes parties d'un corpus. On verra plus loin comment ces tableaux sont utilisés dans différentes analyses statistiques sur les segments répétés.

On convient de classer les segments répétés d'abord par ordre de longueur décroissante (en se bornant en général aux segments de longueur inférieure ou égale à sept) puis, à l'intérieur d'une même classe de longueur, de les classer par ordre de fréquence décroissante ; enfin, si nécessaire, les segments seront départagés par l'ordre lexicographique. Les segments contraints dans des segments plus longs, qui apportent une information redondante, seront écartés de ce type de tableau.

Pour le corpus P le tableau des segments répétés peut donc s'écrire de la manière suivante :

segment \ *partie*	P1	P2	P3
ABC	1	1	0

En effet, tous les autres segments du corpus P sont écartés de ce TSR soit parce qu'ils sont des segments hapax, soit parce qu'ils sont contraints dans le segment ABC.

2.6 Les inventaires de segments répétés (ISR)

Une fois repérés par les procédures informatiques, les segments répétés doivent être répertoriés dans des "inventaires" destinés à être consultés.

Pour classer l'ensemble des segments répétés dans un texte, il est nécessaire de définir un ordre de classement qui permettra de retrouver sans trop de difficultés chacun des segments sans avoir pour cela à parcourir des listes trop volumineuses.
L'expérience montre cependant, que même pour un type de texte défini à l'avance, aucun ordre "canonique" d'édition fixé a priori ne peut prétendre satisfaire l'ensemble des utilisateurs. En effet, comme dans le cas de l'édition d'une concordance et, dans une moindre mesure, de la confection d'index hiérarchiques de formes simples, le mode de classement des unités rassemblées dans les ISR doit satisfaire à la fois à deux types d'exigences.

Tout d'abord, pour permettre une consultation aisée de ces listes, l'ordre qui préside à leur établissement doit s'appuyer sur des caractères formels ne prêtant à aucune ambiguïté (afin de permettre un traitement en machine), faciles à décrire et à mémoriser, pas trop nombreux et, bien évidemment, productifs du point de vue du classement. Pour ces raisons, on retiendra habituellement des critères tels que : la fréquence, l'ordre alphabétique, la longueur, l'ordre d'apparition dans le texte, etc.

Mais le mode de classement des segments a par ailleurs, une autre fonction dont l'importance est loin d'être négligeable. En effet, en combinant et en hiérarchisant, par exemple, lors de l'édition d'une liste, deux ou plusieurs de ces critères formels de classement, on obtient parfois un "effet visuel" qui peut attirer l'attention du chercheur.

Cet "effet" renvoie, selon les cas, à l'existence dans le corpus étudié d'une locution usuelle ou bien encore à l'emploi massif d'une association syntagmatique, dont on n'aurait sans doute pas remarqué l'existence si les lignes de l'inventaire avaient été triées autrement.

Pour l'étude de corpus de textes comptant plusieurs centaines de milliers d'occurrences il est inutile d'imprimer systématiquement des inventaires exhaustifs dont le volume rendrait la consultation peu pratique. Pour mener à bien des recherches de ce type, il est préférable d'automatiser directement la quantification et la recherche des "effets" décrits plus haut. La notion de segment répété dans l'ensemble du corpus se révèle être un outil particulièrement adapté à la mise en évidence de ces récurrences.

Pour toutes ces raisons, il est important de bien maîtriser la combinatoire des modalités de sélection et de classement des segments répétés à partir de critères formels. On détaillera ici plusieurs modes de classement et de sélection en commençant par les plus simples.

2.6.1 Inventaire alphabétique des segments répétés

En consultant l'index alphabétique des formes simples présentes dans un corpus de textes, on est souvent conduit à se demander si certaines de ces formes n'apparaissent pas massivement dans des syntagmes qui les contiennent simultanément.

Tableau 2.9

Extrait d'un inventaire alphabétique des segments répétés réalisé pour le corpus des réponses à la question *Enfants*.

F	L		Segment
14	3		*avenir des enfants*
			318 208 148
2	2		*avenir difficile*
			318 28
4	2	—> 2	*avenir du*
	3		18 155
2	3		*avenir du couple*
			318 155 95
13	2	—> 7	*avenir économique*
			318 61
2	3		*avenir est incertain*
			318 190 45
35	2		*avenir incertain*
			318 45
2	2		*avenir leur*
			318 71
			===> *quel avenir leur . . .*
13	2	—> 7	*avenir pour*
			318 161
7	4		*avenir pour les enfants*
			318 161 442 148
3	5		*avenir qui lui est réservé*
			318 76 15 190 3*

Guide de lecture du tableau 2.9

- La première colonne indique la fréquence du segment dans le corpus.
- Dans la seconde, on lit la longueur du segment (nombre de formes qui le composent).
- On trouve éventuellement dans la troisième colonne, après une flèche orientée vers la droite ---> , la fréquence de l'expansion droite (de longueur immédiatement supérieure) la plus fréquente.
- Sous le segment lui-même, on peut lire les fréquences respectives de chacun des éléments qui le composent. Lorsqu'une forme n'apparaît que dans le contexte constitué par le segment donné, sa fréquence est suivie d'un astérisque.
- Les segments sont parfois munis de références numériques permettant de localiser leurs différentes occurrences dans le corpus.
- Lorsqu'un segment possède une expansion contrainte à gauche (cf. plus haut) une double flèche ====> précède la mention de cette expansion gauche. C'est à cette entrée de l'ISR que l'on trouvera les références du segment contraint omises ici pour cause de redondance.

Ainsi par exemple, dans les réponses aux questions ouvertes étudiées aux chapitres précédents, la présence simultanée des formes *coût* et *vie* induit immédiatement la question de la présence ou non dans le texte étudié du syntagme *coût de la vie* et de sa localisation en tant que tel.

A côté de chacun des segments, on trouve dans l'inventaire des renseignements sur sa fréquence, sa localisation et sur certaines de ses expansions éventuelles.

Si la polyforme que l'on recherche est répétée dans le texte, l'ISR alphabétique permet de répondre à ce genre de question sans trop de difficultés.

On peut voir au tableau 2.9 un extrait de l'ISR alphabétique réalisé à partir des réponses à la question ***Enfants*** qui nous sert d'exemple de référence. Dans ce document, les segments sont triés par ordre lexicographique sur le premier terme; si deux segments ont un terme initial identique, ils sont départagés par l'ordre lexicographique de leur deuxième terme et ainsi de suite[1].

[1] Dans la version des programmes utilisée pour cette expérience, nous nous sommes bornés au recensement des segments répétés composés de sept formes au maximum, afin de ne pas alourdir un algorithme déjà complexe. Les séquences répétées plus longues, en général assez peu nombreuses, sont pour l'instant recensées dans une étape ultérieure. L'expérience montre qu'on peut cependant les repérer assez aisément à partir de leurs sept premières composantes.

Signalons encore que dans l'ISR alphabétique les segments possédant eux-mêmes des expansions droites de même fréquence sont éliminés. En effet dans l'inventaire:

ABC	6 fois
ABCD	4 fois
ABCDE	4 fois

la mention du segment "intermédiaire" ABCD est redondante du fait de la présence à la ligne suivante du segment de même fréquence ABCDE.

2.6.2 Inventaire hiérarchique des segments répétés

L'inventaire alphabétique des segments répétés se révèle peu adapté au repérage des segments les plus longs ou encore des segments les plus fréquemment utilisés dans un corpus.

Tableau 2.10

Extrait d'un inventaire hiérarchique des segments répétés réalisé pour le corpus des réponses à la question *Enfants*.

F	L		Segment
10	5		*le coût de la vie* 474 14 915 666 179
9	5		*le travail de la femme* 474 152 915 666 60
8	5	—> 2	*les difficultés de la vie* 442 82 915 666 179
8	5	—> 3	*peur de ne pas pouvoir* 160 915 193 325 41
7	5	—> 2	*le fait de ne pas* 474 25 915 193 325
6	5		*la cherté de la vie* 666 7 915 666 179

Dans l'inventaire hiérarchique, les segments sont triés en premier lieu par ordre de longueur décroissante, les segments de même longueur étant rangés par ordre de fréquence décroissante. Enfin, l'ordre lexicographique départage

les segments qui seraient encore en concurrence à ce niveau. On peut voir au tableau 2.10 un fragment de l'ISR hiérarchique du corpus des réponses à la question ***Enfants*** correspondant aux segments de longueur 5. Les informations qui concernent chacun des segments sont identiques à celles décrites à propos des inventaires alphabétiques[1].

Inventaire hiérarchique par partie

Il est utile de disposer, pour chacune des parties du corpus, d'un inventaire hiérarchique des segments répétés ne prenant en compte que les occurrences des segments répétés dans cette seule partie.

Tableau 2.11
Inventaire distributionnel avant la forme : *enfants*
(Question *Enfants*)

les enfants	----	----	----	----	----	59
pour les enfants	----	----	----	----	22	
avenir pour les enfants	----	----	----	7		
chômage pour les enfants	----	----	----	4		
du chômage pour les enfants	----	----	2			
travail pour les enfants	----	----	----	2		
aimer les enfants	----	----	----	----	7	
fait de ne pas aimer les enfants	----	----	----	2		
pas les enfants	----	----	----	----	7	
aiment pas les enfants	----	----	----	5		
aime pas les enfants	----	----	----	2		
que les enfants	----	----	----	----	4	
élever les enfants	----	----	----	----	3	
embêter avec les enfants	----	----	----	----	2	
des enfants	----	----	----	----	----	55
avenir des enfants	----	----	----	----	14	
avoir des enfants	----	----	----	----	9	
pour avoir des enfants	----	----	----	---2		
occuper des enfants	----	----	----	----	3	
vis a vis des enfants	----	----	----	----	3	
éduquer des enfants	----	----	----	----	2	
élever des enfants	----	----	----	----	2	
faire des enfants	----	----	----	----	2	
la garde des enfants	----	----	----	----	2	
de mettre des enfants	----	----	----	----	2	
par des enfants	----	----	----	----	2	
la situation des enfants	----	----	----	----	2	
sans enfants	----	----	----	----	----	6
ses enfants	----	----	----	----	----	6
peur du chômage pour ses enfants	----	----	----	----	2	
aux enfants	----	----	----	----	----	2

[1] Il faut noter que l'on ne peut, dans le cas des inventaires hiérarchiques, éliminer les informations redondantes selon les critères retenus pour la confection des inventaires alphabétiques. Voisins dans les inventaires alphabétiques, les segments répétés et leurs expansions récurrentes, droites ou gauches, se trouvent dispersés dans les inventaires hiérarchiques, en raison de leurs longueurs respectives différentes.

Comme dans l'ISR hiérarchique total, les segments sont d'abord triés en fonction de leur longueur. Les segments de même longueur sont ensuite triés d'après leur fréquence dans la seule partie concernée par l'inventaire.

L'ordre alphabétique départage les segments de même longueur qui ont une fréquence égale.

2.6.3 Inventaires distributionnels des segments répétés

Les inventaires distributionnels réalisés à partir d'une forme rassemblent les expansions récurrentes de cette forme. On trouve au tableau 2.12 un inventaire distributionnel[1] réalisé pour les expansions gauches de la forme *enfants*. L'inventaire s'ouvre sur l'expansion de longueur 2 la plus fréquente *les enfants* qui compte 59 occurrences. Les lignes qui suivent sont constituées par les expansions de longueur 3 de cette même polyforme.

Comme on peut le vérifier, pour chaque polyforme de longueur donnée, les expansions de longueur immédiatement supérieure sont classées par ordre de fréquence décroissante. Les expansions de même longueur n et de même fréquence sont départagées par la fréquence de l'expansion gauche de longueur n+1.

Lorsqu'une polyforme possède une expansion de longueur 1 contrainte à gauche (dans le texte pour lequel a été réalisé l'inventaire) elle apparaît précédée de cette expansion. Sur la gauche du tableau on trouve les fréquences de chacune des formes qui composent les polyformes répétées.[2]

Symétriquement, on peut construire un inventaire distributionnel qui classe, selon les mêmes principes, les expansions récurrentes situées à droite d'une forme-pôle. Cet inventaire concerne en général des expansions différentes qui éclairent sous un autre jour l'utilisation de la forme dans le corpus.

2.6.4 Tableau des segments répétés

Le tableau des segments répétés (TSR) permet de juger de la ventilation de chacun des segments dans les différentes parties du corpus. L'ordre des segments dans ce tableau est le même que celui utilisé dans l'inventaire hiérarchique. C'est ce tableau qui sert de base à la plupart des études

---

[1] Cette forme d'inventaire est issue d'un travail en collaboration avec P. Fiala.

[2] On notera qu'à la différence des concordances, les inventaires distributionnels des expansions récurrentes ne fournissent pas d'information sur les expansions non-répétées. Ainsi, par exemple, l'inventaire présenté au tableau 3.12 ne concerne que 132 des 148 occurrences de la forme *enfants* dans le corpus que nous étudions.

statistiques faites à partir des segments. Les segments contraints dans des segments plus longs sont écartés de ce tableau.

Tableau 2.12

Ventilation de quelques segments de longueur 4 dans les parties (Question ouverte *Enfants*)

F	L	Diplôme / Age	A -30	A -50	A +50	B -30	B -50	B +50	S -30	S -50	S +50
33	4	la peur de l	2	7	6	6	3	1	2	5	1
20	4	je ne sais pas	1	2	13	1	2	0	1	0	0
15	4	les conditions de vie	0	1	7	2	0	1	2	0	2
12	4	coût de la vie	3	2	3	2	0	1	0	0	1
12	4	le manque de travail	2	4	5	1	0	0	0	0	0
12	4	peur de ne pas	1	1	3	2	1	0	1	0	3
12	4	travail de la femme	1	0	3	2	2	1	0	2	1
11	4	difficultés de la vie	0	3	2	1	1	2	0	1	1
10	4	le coût de la	2	2	2	2	0	1	0	0	1
9	4	de ne pas pouvoir	1	0	2	1	1	0	1	1	2
8	4	de plus en plus	0	5	2	0	0	0	1	0	0

Le tableau 2.12 donne un extrait du TSR du corpus correspondant aux segments répétés de longueur 4. La première colonne donne un numéro d'ordre qui permet de retrouver le segment dans les analyses statistiques, on trouve ensuite le segment répété puis la fréquence du segment dans le corpus. Les dernières colonnes donnent les sous-fréquences du segment dans chacune des parties du corpus.

2.7 Recherche de cooccurrences, quasi-segments

Les recherches sur les segments répétés se sont développées à partir des difficultés rencontrées dans le domaine de la recherche des cooccurrences (attirances particulières entre couples de formes au sein d'une unité de contexte donnée). Sans prétendre à l'exhaustivité, car les méthodes utilisées dans ce but varient fortement en fonction des domaines d'application, on mentionnera très brièvement, dans les paragraphes qui suivent, quelques-unes des méthodes élaborées pour répondre à cette préoccupation.

2.7.1 Recherche autour d'une forme-pôle

Pour une forme pôle donnée, plusieurs méthodes permettent de sélectionner un ensemble de formes qui ont tendance à se trouver souvent dans un voisinage de cette forme.

Pour sélectionner ces formes il faut commencer par définir une unité de contexte, ou voisinage, à l'intérieur duquel on considérera que deux formes sont cooccurrentes. Cette unité peut ressembler à la phrase[1] ou encore être constituée par un contexte de longueur fixe (x occurrences avant, et x occurrences après la forme-pôle).

Labbé (1990) propose une méthode particulièrement simple destinée à mettre en évidence ce qu'il appelle l'*univers lexical* d'une forme donnée. Pour chaque forme *forme1* du corpus, l'ensemble des phrases du corpus peut être divisé en deux sous-ensembles : P_1, sous-ensemble de celles qui contiennent *forme1* et P_0, sous-ensemble des unités desquelles *form1* est absente.

Pour chacune des autres formes du corpus, on applique ensuite le test de l'écart-réduit aux sous-fréquences dans les deux ensembles P_0 et P_1 en tenant compte de leurs longueurs respectives. Si les fréquences des formes considérées ne sont pas trop faibles cette méthode permet de sélectionner, pour chaque forme pôle donnée un ensemble de formes qui se trouvent situées de manière privilégiée dans les mêmes phrases.

La *méthode des cooccurrences n°1*, proposée par Lafon et Tournier (1970)[2] différait de la méthode exposée plus haut sur deux points principaux :

a) elle distinguait les positions relatives par rapport à la forme-pôle séparant de ce fait des cooccurrences *avant* et des cooccurrences *après*.

b) elle construisait un coefficient de cooccurrence faisant intervenir à la fois la cofréquence des deux formes et leur distance moyenne mesurée en nombre d'occurrences.

La *méthode des cooccurrences n°3*, Lafon (1981) affecte un indice de probabilité au nombre des rencontres, chaque fois à l'intérieur d'une phrase (séquence de formes entre deux ponctuations fortes), de chaque couple et de chaque paire (couple non-orienté) de formes du texte. Cette méthode permet de sélectionner, à l'aide d'un seuil en probabilité des ensembles de couples et

[1] Si l'on accepte de se satisfaire de critères relativement simples permettant de repérer les limites des phrases dans la plupart des cas, mais cette question est plus difficile qu'elle ne le semble au premier abord.

[2] Cf. Demonet et al. (1975)

de paires de formes présentant des affinités dans le corpus de textes que l'on étudie. Tournier (1985b) a appliqué ce type de recherche de cooccurrences à la construction de *lexicogrammes* (réseaux orientés de formes cooccurrentes).

D'une manière générale, la recherche d'associations privilégiées est un enjeu important dans les applications relevant de l'industrie de la langue. Ainsi, certaines incertitudes rencontrées lors de la lecture optique de caractères peuvent être levées (au moins en probabilité) par la considération des formes voisines déjà reconnues, si l'on connaît les probabilités d'association. La désambiguïsation lors d'une analyse morpho-syntaxique peut être réalisée dans les mêmes conditions.

Les travaux de Church et Hanks (1990) se situent dans ce cadre général. Ces auteurs proposent d'utiliser comme mesure d'association entre deux formes *x* et *y* l'information mutuelle *I(x, y)*, issue de la théorie de la communication de R. Shannon :

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

où *P(x)* et *P(y)* sont les fréquences des formes *x* et *y* dans un corpus, et *P(x,y)* la fréquence des occurrences voisines de ces deux formes, x précédant y (il n'y a donc pas symétrie vis-à-vis de x et y), le voisinage étant défini par une distance comptée en nombre de formes. Ainsi, pour les textes en anglais, ces auteurs préconisent de considérer comme voisines deux formes séparées par moins de cinq formes.[1]

2.7.2 Recherches de cooccurrences multiples

On a rangé sous ce chapitre des méthodes qui se fixent pour objet le repérage direct des cooccurrences de deux ou plusieurs formes du texte. Ce dernier type de recherche s'est tout particulièrement développé dans le domaine de la recherche documentaire.

Dans le domaine de l'indexation automatique on utilise des coefficients qui permettent de repérer des cooccurrences de plusieurs termes à la fois Ainsi, par exemple, Chartron (1988) a proposé un coefficient dit *coefficient d'implication réciproque*.

---

[1] Cette procédure paraît efficace pour identifier certaines séquences "verbe + préposition" (*phrasal verbs*). Sur un corpus de 44 millions d'occurrence, les auteurs montrent ainsi que *set* est suivi d'abord de *up (I(x,y)* = 7.3), puis de *off* (*I(x,y)* = 6.2), puis de *out, on in,* etc.

Pour un groupe composé de k formes numérotées 1, 2, 3, ... , k on calcule le coefficient E de la manière suivante :

$$E = \frac{N_{12..k}}{N_1 N_2 \ldots N_k}$$

avec : $N_{12...k}$ = nombre de cooccurrences des k formes dans le corpus
$N_1, \ldots, N_k$ = nombre d'occurrences de chacune des k formes

D'autres chercheurs ont proposé des approches de ce même problème à partir de la notion de représentation arborescente.[1]

2.7.3 Quasi-segments

Les procédures qui sélectionnent les segments répétés dans un corpus ne permettent pas de repérer, dans l'état actuel, des répétitions légèrement altérées, par l'introduction d'un adjectif ou par une modification lexicale mineure touchant l'un des composants. Bécue (1993) a proposé un algorithme qui permet de repérer des *quasi-segments* (répétés). Cet algorithme permet, par exemple, de rassembler en une même unité *(faire, sport)* les séquences *faire du sport* et *faire un peu de sport*. Il s'agit d'une direction de recherche prometteuse, mais les quasi-segments sont encore plus nombreux que les segments, et leur recensement pose des problèmes de sélection et d'édition.

2.8 Incidence d'une lemmatisation sur les comptages

Pour clore ce chapitre, nous avons choisi de présenter une comparaison de comptages, effectués à la fois en termes de formes graphiques et en termes d'unités lemmatisées, sur un corpus relativement important du point de vue de sa taille.

Cette comparaison permettra de mesurer l'incidence effective du travail de lemmatisation sur les décomptes lexicométriques. Nous analyserons plus loin, (Chapitre 7, paragraphe 7.6) les conséquences de cette modification de l'unité de décompte sur les typologies portant sur l'ensemble des parties de ce même corpus.

[1] Cf. par exemple : Barthélémy et al. (1987) et Luong X.(1989).

2.8.1 Le corpus *Discours*

Le corpus que l'on a retenu rassemble les textes de 68 interventions radiotélévisées de F. Mitterrand survenues entre juillet 1981 et mars 1988.

Tableau 2.13

Les principaux codes grammaticaux utilisés pour la lemmatisation du corpus *Discours*

Codes grammaticaux

1 Verbe :
— 11 *forme fléchie*
— 12 *forme au participe passé*
— 13 *forme au participe présent*
— 14 *forme à l'infinitif*

2 Substantif :
— 21 *substantif masculin*
— 22 *substantif féminin*
— 23 *substantif homographe masculin*
— 24 *substantif homographe féminin*

3 Adjectif :
5 Pronom
6 Adverbe
7 Déterminant
8 Conjonction
9 Locution

P Ponctuation
— p *ponctuation mineure (interne à la phrase)*
— P *ponctuation majeure (délimitant la phrase)*

A partir des textes de ces interventions retranscrits sur support magnétique, Dominique Labbé[1] a effectué un travail de lemmatisation selon des normes

[1] L'ensemble des données textuelles que nous avons utilisées dans ce paragraphe a été rassemblé par D. Labbé, chercheur au *Centre de Recherche sur le Politique, l'Administration et le Territoire (CERAT -Grenoble)*, pour une étude sur le vocabulaire de F. Mitterrand. Cet ensemble est constitué à la fois par le texte de discours prononcés au cours du premier septennat et un fichier de lemmatisation qui indique pour chacune des occurrences du texte un lemme de rattachement et une catégorisation grammaticale de l'occurrence considérée. D. Labbé nous a rendu accessible l'ensemble des données dont il

qu'il décrit dans Labbé (1990b). Un patient travail partiellement assisté par ordinateur, lui a permis de rattacher chacune des formes graphiques du texte d'origine à un lemme.

On trouve au tableau 2.14 un exemple de fichier-résultat obtenu à la suite d'un tel processus. Chaque ligne du fichier-résultat correspond à une occurrence graphique ou à une ponctuation du texte d'origine.

— la première colonne contient la forme graphique attestée dans le texte;
— la seconde le lemme auquel le processus de lemmatisation rattache cette forme graphique;
— la dernière colonne donne la catégorie grammaticale du lemme.

Il est alors possible de reconstituer la séquence de formes lemmatisées qui va être soumise aux analyses statistiques.

Tableau 2.14

Exemple de fichier de lemmatisation

Forme	*Lemme*	*Cat.Gr.*	*Forme*	*Lemme*	*Cat.Gr.*
je	je	5	le	le	7
crois	croire	11	quatorze	quatorze	7
qu'	que	82	juillet	juillet	21
on	on	5	c'	ce	5
ne	ne	6	est	être	11
peut	pouvoir	11	sans	sans	81
que	que	82	aucun	aucun	7
souhaiter	souhaiter	14	doute	doute	21
cela	cela	5	-	-	p
/.	/.	P			

Le caractère $ sera utilisé pour séparer la suite de caractères qui indique le lemme de rattachement de la partie numérique, laquelle indique une catégorie grammaticale.

A partir du texte ainsi transcodé, il va être possible, en utilisant les mêmes algorithmes que ceux que nous avons utilisé pour effectuer les décomptes de formes graphiques, d'obtenir des comptages portant cette fois sur les lemmes ainsi codés.

---

disposait afin de nous permettre de faire les quelques expériences dont nous rendons compte ici. D. Labbé a publié plusieurs ouvrages portant sur le discours politique. Deux de ces ouvrages, Labbé (1983) et Labbé (1990) portent plus particulièrement sur le discours de F. Mitterrand.

Tableau 2.15

Discours - Extrait du discours de F. Mitterrand(14/07/81) avant et après l'opération de lemmatisation

je crois qu'on ne peut que souhaiter cela. le 14 juillet c'est sans aucun doute - et c'est fort important - l'occasion d'une revue, d'un défilé, d'une relation directe entre notre armée et la nation.

_____

je$5 croire$11 que$82 on$5 ne$6 pouvoir$11 que$82 souhaiter$14 cela$5. le$7 quatorze$7 juillet$21 ce$5 être$11 sans$81 aucun$7 doute$21 - et$82 ce$5 être$11 fort$6 important$3 - le$7 occasion$22 de$81 un$7 revue$22, de$81 un$7 défilé$21, de$81 un$7 relation$22 direct$3 entre$81 notre$7 armée$22 et$82 le$7 nation$22

2.8.2 Comparaison des principales caractéristiques quantitatives

Le tableau 2.16 permet une première comparaison entre les décomptes effectués sur les lemmes et ceux réalisés à partir des formes graphiques du corpus.

Tableau 2.16

Principales caractéristiques lexicométriques des dépouillements en formes graphiques et en lemmes

	formes	lemmes
nombre des occurrences :	297258	307865
nombre des formes :	13590	9309
nombre des hapax :	5543	3255
fréquence maximale :	11544	29559

Comme on peut le vérifier sur ce tableau, le nombre des occurrences est plus élevé pour le corpus lemmatisé (+10 607 occurrences). Cette circonstance n'étonnera pas outre mesure car deux grands types d'opérations influent de manière inverse sur le nombre des occurrences d'un corpus dans le cas d'une lemmatisation du type de celle qui a été considérée plus haut.

— le repérage de certaines unités composées de plusieurs formes graphiques (*à l'instar, à l'envi, d'abord, d'ailleurs, etc.*) tend à réduire le nombre des occurrences du corpus lemmatisé.

— à l'inverse, l'éclatement en plusieurs unités distinctes de chacune des nombreuses occurrences des formes graphiques contractées (*au = à + le, des = de + les, etc.*.) tend pour sa part à augmenter le nombre des occurrences du fichier lemmatisé par rapport au texte initial.

L'excès important du nombre des *occurrences* constaté pour le fichier lemmatisé par rapport au texte d'origine tend simplement à prouver, comme on pouvait d'ailleurs s'y attendre, que le second des phénomènes évoqués ci-dessus est nettement plus lourd que le premier.

On peut expliquer par des raisons analogues le fait que le nombre total des *formes* du corpus lemmatisé est, quant à lui, nettement inférieur au nombre des formes graphiques différentes dans le corpus non-lemmatisé. Ici encore deux phénomènes jouent en sens contraire lors du passage aux unités lemmatisées.

— le regroupement de plusieurs formes graphiques correspondant aux différentes flexions d'un même lemme (flexions verbales, marques du genre et du nombre pour les adjectifs, etc.) réduit le nombre des formes du corpus lemmatisé.

— la désambiguïsation rendue possible par la détermination de la catégorie grammaticale pour chacune des formes (*le*-pronom vs *le*-article*, etc.*) augmente plutôt ce nombre.

Au plan quantitatif, on le voit, c'est le processus de regroupement des formes graphiques qui l'emporte nettement lors d'une lemmatisation.

Ces conclusions valent également, nous semble-t-il, pour le nombre des hapax du corpus lemmatisé nettement inférieur lui aussi à son homologue mesuré en formes graphiques.

La fréquence maximale du corpus lemmatisé correspond aux occurrences du lemme *le-article* qui regroupe celles des occurrences des formes graphiques : *le, la, les, l'* qui ne correspondent pas à des occurrences des pronoms homographes[1]. Cette fréquence se révèle plus de 2 fois supérieure à son homologue calculée à partir du corpus non-lemmatisé.

[1] On se souvient que dans la plupart des corpus de textes écrits en français que nous avons dépouillés en décomptant les formes graphiques, la forme la plus fréquente est la forme *de*. Tel est également le cas pour le corpus ***Discours***.

Mesure de l'accroissement du vocabulaire

La figure 2.5 rend compte de l'accroissement du vocabulaire, mesuré à l'aide des deux types d'unités. Comme on le voit, l'allure générale de la courbe d'accroissement rappelle bien dans les deux cas les courbes qui ont été observées pour les corpus dépouillés précédemment. Cependant, le nombre des formes nettement inférieur dans le cas du corpus lemmatisé fait que les deux courbes paraissent assez différentes au premier abord.

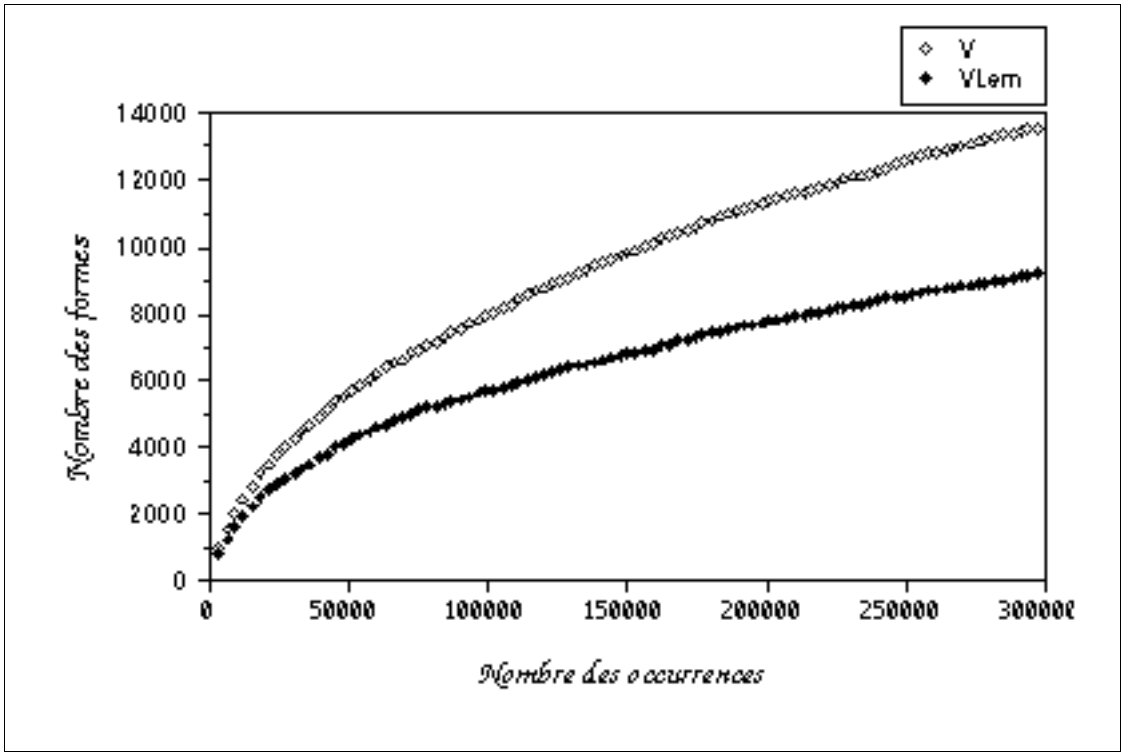

Figure 2.5

L'accroissement du vocabulaire
mesuré en formes graphiques et en lemmes.

Ces circonstances incitent à limiter les comparaisons, en ce qui concerne les principales caractéristiques lexicométriques, à des comptages réalisés selon des normes de dépouillement identiques.

On verra cependant au chapitre 7, à propos des séries textuelles chronologiques (paragraphe 7.6) que les typologies réalisées sur une même partition du corpus présentent une grande stabilité que l'on retienne l'une ou l'autre des normes de dépouillement du texte.