CHAPITRE IX

QUANTIFIER LES FAITS LANGAGIERS

Divers outils informatiques permettent d'extraire, à partir de corpus ayant fait l'objet d'un travail d'annotation, les occurrences d'unités textuelles qui correspondent à un patron donné (mot, lemme, catégorie grammaticale ou sémantique, patron syntaxique, etc.). Ces outils permettent aisément de constituer la liste exhaustive des contextes où cette unité-pôle apparaît. L'examen des différents contextes d'une unité textuelle projette un éclairage indispensable sur les emplois que cette unité trouve dans le corpus, faisant apparaître des régularités qu'une lecture cursive du corpus n'aurait pas toujours révélées. Cependant dès que le nombre des contextes est un peu élevé, les mises en contextes ainsi réalisées (comme les concordances, etc.) deviennent des objets difficilement manipulables, même sous forme informatisée. L'organisation de ces listes (définition et ordre de présentation des contextes) influence très fortement la perception de divers phénomènes relatifs à la forme-pôle.

Le tableau 1 regroupe quelques lignes extraites des 5 030 contextes de la forme *je* dans ***Mitterrand1***. Ces contextes sont triés par ordre alphabétique, d'après la forme qui suit le pôle. Une telle approche permet de remarquer, en inspectant l'ensemble des lignes de contexte réalisées pour cette forme, que les occurrences de *je* sont prises dans des répétitions plus longues: *je le crois, je le dis*, etc.

Tableau 1. — Extrait d'une concordance de la forme je dans **Mitterrand1**

ue la france qui a acquis,	je le crois,	la confiance et le respect
ères personnels, aussi, et	je le crois,	qui se réfèrent à la moral
cer des propositions pour,	je le crois,	saisir le monde entier du
rté des facilités qui ont,	je le crois,	sauvé le secteur du textil
ation de la fin du siècle.	je le crois	tout à fait, sans quoi je n
n souvient aussi- cela est	je le crois,	tout à fait, venu de consi
de la république: je suis,	je le crois,	très fidèle à ce que je su
jours, j' ai observé avec,	je le crois,	une grande patience, pour
ants que cela contribuera,	je le crois,	utilement au redressement
bre de plans, j' ai donné-	je le crois	vraiment- plus d' expansion
rachever le portrait. moi,	je le dessine	tous les jours, par des a
ite, je l' ai dit à alger,	je le dirai	à amman en jordanie où je s
dans le monde. la france,	je le dirai	simplement, a déjà apporté

Pour généraliser ce type de démarche à l'ensemble des formes du corpus, il faut mettre en oeuvre des procédures de quantification qui éviteront au chercheur

d'avoir à examiner l'ensemble des contextes de chacune des formes du corpus.

Ce chapitre propose un survol des approches quantitatives les plus courantes d'un corpus de textes[1]. La section 1 présente des objectifs de recherche qui conduisent à opérer des décomptes textuels à des fins de comparaison. Les problèmes liés à l'identification des unités dans le texte sont abordés dans la section 2. La section 3 traite du repérage des séquences d'unités. Les sections 4 et 5 introduisent ensuite des méthodes permettant de comparer les décomptes réalisés au sein d'un corpus partitionné. La section 6 est consacrée à l'articulation des décomptes réalisés à partir de différents systèmes d'annotation. Nous terminons (section 7) par un exemple de recherche sur les séries textuelles chronologiques qui combine plusieurs des méthodes présentées dans le chapitre.

Pourquoi quantifier ?

Au-delà des études centrées chaque fois sur un type d'unité textuelle particulier, s'est développé un courant dont les dénominations ont varié au cours du temps[2], et qui se fixe pour but l'étude quantitative des faits langagiers. L'approche quantitative permet seule d'accéder à la description de phénomènes textuels qui présentent un grand intérêt une fois mis en évidence et dont il aurait été difficile de cerner les contours *a priori*..

Étudier la variation de traits linguistiques dans un corpus

Certaines études menées par des linguistes se fixent pour but principal la description de la variation, au sein d'un corpus, de l'ensemble des éléments d'un même système d'unités linguistiques (graphèmes, formes, lemmes, lexies, système de catégories grammaticales, séquences, etc.). En général, ce type de tâche s'accommode mal de procédures de segmentation et d'identification approximatives des unités de décompte. Il nécessite au contraire que le texte analysé soit soumis, lors d'une étape préalable, à une réflexion minutieuse sur les procédures de repérage, d'identification et d'annotation des unités à recenser. Une fois les comptages réalisés pour chacune des unités du système, on soumet ces décomptes à des traitements statistiques afin de mettre en évidence les variations des différentes unités.

Réaliser des typologies de textes et de documents

Un courant relativement ancien de l'analyse quantitative des textes opère des quantifications dans le but de réaliser des typologies portant sur l'ensemble des textes réunis en corpus. Le problème de l'attribution d'auteur[3] en est un exemple. Il s'agit de déterminer si tel ou tel texte, sur lequel on manque de renseignements, présente des caractéristiques quantitatives laissant supposer qu'il a pu être écrit par un auteur dont on possède par ailleurs des échantillons de textes. On s'efforce donc de déterminer des systèmes d'unités discriminantes qui permettent de trancher en matière d'attribution. La comparaison des descriptions quantitatives des différents textes doit permettre dans ce cas d'obtenir des indications qui ne résultent pas de connaissances *a priori* sur les textes mais bien des similitudes qu'ils présentent au plan quantitatif.

---

[1] Chacune de ces méthodes est présentée dans (Lebart et Salem, 1994).

[2] Cf., par exemple, (Herdan, 1964), (Muller, 1968).

[3] Le travail de (Holmes, 1985) présente une revue assez complète des travaux en matière d'attribution d'auteur.

On a recours à des méthodes comparables lorsqu'il s'agit de prélever parmi un vaste ensemble de documents ceux d'entre eux qui peuvent présenter de l'intérêt pour une tâche particulière (problème de l'indexation et de la récupération de documents industriels).

Pour ce second type d'études, le problème de la nature linguistique des unités qui permettent de mener à bien les tâches entreprises n'est pas central puisque le but ultime est le regroupement de textes. La sélection du système des unités de décompte qui sert de base aux comparaison se fait avant tout en fonction de l'efficacité pratique de l'ensemble de la démarche au regard de la tâche considérée.

Ces deux types de préoccupation (sections 1.2 et 1.2) se combinent parfois en proportions variables dans des études particulières. La mise en place de procédures à visées typologiques pose du même coup le problème du choix des unités les mieux à même de faire ressortir des oppositions.

Déceler des corrélations entre phénomènes

Une étude portant sur la répartition des pronoms personnels de la première personne dans chacune des huit années de ***Mitterrand1*** montre que la fréquence d'emploi de ces pronoms varie sensiblement au cours du temps. On constate sur la figure 1, une tendance à l'augmentation du pronom *je* et une diminution du pronom *nous*. Cette tendance s'inverse légèrement dans la dernière année du septennat. Comme on le voit, les deux phénomènes manifestent une certaine liaison au cours du temps.

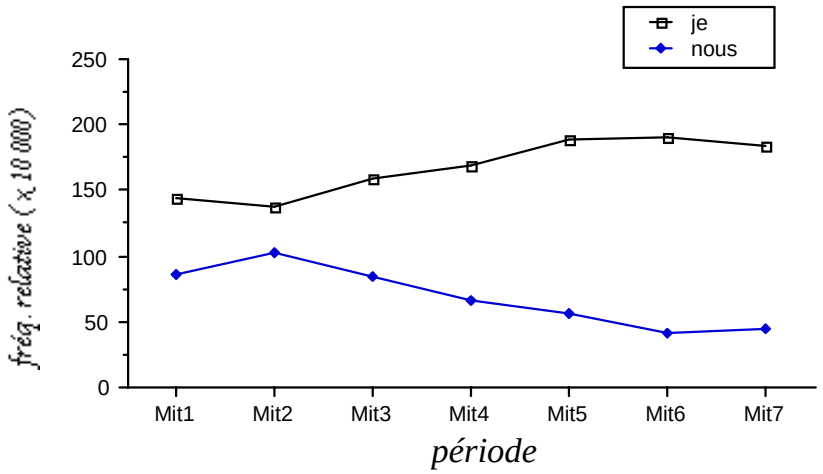


Figure 1.— Évolution des formes je et nous dans Mitterrand1[4]

On comprend aisément que ces variations de fréquences intéressent des spécialistes du texte politique. L'entrée quantitative est ici la seule voie d'accès à l'analyse détaillée et contrastive d'un tel phénomène.

Les unités

La méthode statistique s'appuie sur des mesures et des comptages réalisés à partir des objets que l'on veut étudier. Décompter des unités, les additionner entre elles,

[4] Le nombre des occurrences de chaque forme, dans chaque partie, est rapporté à la longueur de la partie considérée et multiplié par 10 000 pour une plus grande lisibilité des résultats.

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. 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 de dépouillement* permettant d'isoler à partir du texte annoté les différentes unités sur lesquelles porteront les dénombrements.

Ch. Muller (1973) 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 [...] et paradigmatique [...] 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 a été rendu discontinu [...].

Normes de dépouillement

Malgré les connotations véhiculés par le mot *norme* dans le domaine linguistique, la notion de *norme de dépouillement* doit être ici comprise comme une exigence de standardisation provisoire des textes contenus dans un corpus. Cette standardisation est destinée avant tout à les rendre comparables, à les stabiliser le temps d'une expérience.

Nous allons illustrer sur un court extrait de ***Mitterrand1***, les problèmes liés à l'établissement d'une telle norme. Le premier fragment de texte (état A) correspond au texte tel qu'il a été saisi au départ.

État A : Texte de départ

> 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.

Le second (Norme B) montre le même extrait du corpus après quelques transformations de surface destinées à permettre l'identification automatique des mêmes formes indépendamment de leur position dans la phrase (les majuscules de début de phrase ont été transformées en minuscules). Les barres verticales matérialisent la segmentation des unités.

Norme B : Elimination des majuscules de début de phrase

> **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.

Dans une phase suivante (Norme C), on a réuni certaines unités polylexicales.

Norme C : Regroupement d'unités polylexicales

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 | . |

Dans les deux états suivants, les mots du texte ont été remplacés par des étiquettes (respectivement : des lemmes – Norme D – et des catégories grammaticales – Norme E).

Norme D : Lemmatisation

je | croire | que | on | ne | pouvoir | que | souhaiter | cela | . | le | quatorze | juillet | ce | être | sans | aucun | doute | - | et | ce | être | fort | important | - | le | occasion | de | un | revue | , | de | un | défilé, | de | un | relation | direct | entre | notre | armée | et | le | nation | .

Norme E : Catégorisation en parties du discours

{pronom} | {verbe} | {subordonnant} | {pronom} | {adverbe} | {verbe} | {subordonnant} | {verbe} | {pronom} | {ponctuation} | {déterminant} | {numéral} | {nom} | {pronom} | {verbe} | {préposition} | {déterminant} | {nom} | {ponctuation} | {coordonnant} | {pronom} | {verbe} | {adverbe} | {adjectif} | {ponctuation} | {déterminant} | {nom} | {préposition} | {déterminant} | {nom} | {ponctuation} | {préposition} | {déterminant} | {nom} | {ponctuation} | {préposition} | {déterminant} | {nom} | {adjectif} | {préposition} | {déterminant} | {nom} | {coordonnant} | {déterminant} | {nom} | {ponctuation}

Le dernier état du texte résulte d'un étiquetage permettant d'identifier les occurrences de quelques indices énonciatifs.

Norme F : Repérage d’indices énonciatifs

{embrayeur} {non-personne} {non-personne} {non-personne} {non-personne} {embrayeur}

Remarquons que, dans le cas de la mise en oeuvre de cette dernière norme de dépouillement, il ne s'agit plus d'une segmentation du texte de départ**.**

Décomptes automatisés

A la phase de délimitation des unités (qui peut être une segmentation) succède une phase de regroupement de celles que l'on considère comme identiques le temps de l'expérience (identification).

Pour un même texte, les différentes normes de dépouillement ne conduisent pas aux mêmes décomptes. Dans chaque expérience pratiquée, ces normes ne présentent pas le même degré de pertinence, ni les mêmes avantages (ou inconvénients) quant à leur mise en oeuvre. Néanmoins, au-delà des

considérations propres à chaque domaine, une fois définie la norme de dépouillement et sa jurisprudence, les méthodes de la *statistique* s'appliquent de manière aveugle aux décomptes réalisés à partir de chacune des normes.

Comme on peut le voir sur les index réalisés à partir de ces transformations du texte de départ, le système des fréquences des unités soumises aux décomptes dépend étroitement de la norme de dépouillement retenue.

On voit sur ce petit exemple la grande latitude des choix possibles quand aux types de décomptes que l'on peut opérer à partir d'un même texte muni d'annotations. Pour chaque recherche particulière, ces choix résultent avant tout des objectifs de recherche poursuivis.

Norme A		**Norme B**		**Norme E**		**Norme F**	
,	4	,	4	{préposition}	15	{non-personne}	4
d’	3	d’	3	{déterminant}	8	{embrayeur}	2
c’	2	-	2	{nom}	8		
est	2	*c’_est*	2	{ponctuation}	6		
et	2	.	2	{pronom}	5		
une	2	et	2	{verbe}	5		
14	1	une	2	{adverbe}	2		
armée	1	nation	1	{coordonnant}	2		
aucun	1	ne	1	{subordonnant}	2		
cela	1	notre	1	{adjectif}	2		
.....		.....		{numéral}	1		
34 types		31 types		11 types		2 types	
45 occ.		40 occ.		56 occ.		6 occ.	

Incidence de la norme sur les décomptes

Mitterrand1 a été soumis à des dépouillements prenant en compte les différents systèmes d'unités évoqués plus haut. On a utilisé successivement :

- le système des caractères qui servent à encoder le texte sur support magnétique ;
- la segmentation du texte en formes graphiques obtenue en déterminant un ensemble de caractères délimiteurs (le point, la virgule, le point et virgule, etc.) ;
- la segmentation du texte en « lemmes » obtenue selon un ensemble de règles fixées par (Labbé, 1995) ;
- un système d'annotations grammaticales comportant 15 catégories différentes (nom, verbe, etc.) élaborée dans le cadre de cette même étude.

Le tableau 2 permet une comparaison rapide entre ces différents décomptes effectués à partir de niveaux d'annotation différents.

Tableau 2.— Décomptes sur Mitterrand1[5]

	caractères	formes	lemmes	catégories
nombre des occurrences :	1 667 251	297 258	307 865	307 865
nombre des types :	98	13 590	9 309	15
nombre des hapax[6] :	0	5 543	3 255	0
fréquence maximale :	224 865*	11 544	29 559	86 700 *

[5] Les décomptes suivi de l'astérisque résultent d'une approximation statistique.

[6] Du grec *hapax legomenon* : chose dite une fois.

Les différents systèmes de décomptes produisent des descriptions difficilement comparables. Le système des catégories compte en effet un nombre relativement faible de types différents, les deux systèmes de descripteurs « lexicaux » (formes et lemmes) ont en commun de posséder un nombre très élevé de types s'étalant sur une large gamme de fréquence.

Exemple : l'accroissement du vocabulaire

Le problème de l'accroissement du vocabulaire (apparition de formes nouvelles au fur et à mesure que l'on avance dans la lecture du corpus) a été largement étudié dans les travaux de la statistique textuelle. La figure 2 rend compte de l'accroissement du vocabulaire, mesuré en lemmes et en formes graphiques. Les deux courbes ont la même *allure générale*. À un accroissement relativement fort au début du corpus, succèdent des périodes d'accroissement plus modestes, bien que tout allongement du corpus entraîne toujours l'apparition de nouvelles formes. Le nombre de formes nettement inférieur dans le cas du corpus lemmatisé fait que la deuxième courbe est toujours largement située en dessous de la première. En fait, deux tendances contraires influent sur les rapports qu'entretiennent ces nombres :

- 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 corpus lemmatisé par rapport au texte initial.

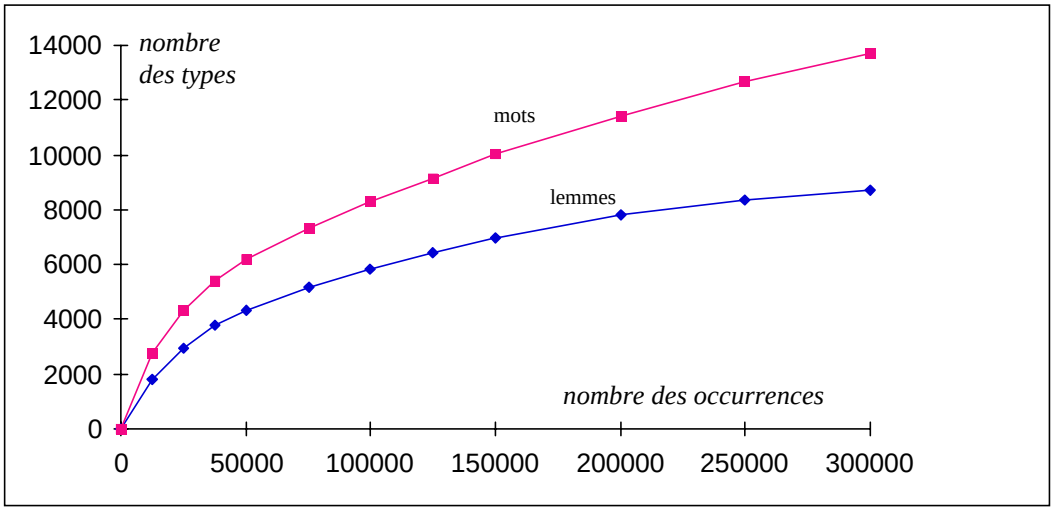


Figure 2. — L'accroissement du vocabulaire mesuré en formes graphiques et en lemmes

Cet exemple souligne la nécessité de pratiquer des comparaisons sur des comptages réalisés selon des normes de dépouillement identiques.

Mesures de récurrence sur l'axe syntagmatique

Les opérations de comptage des unités dans un corpus passent nécessairement par une phase de délimitation qui isole ces dernières de leur contexte immédiat. L'expérience montre cependant qu'après cette phase préliminaire, il est intéressant d'étudier en outre les récurrences et cooccurrences d'unités composées (suite de catégories syntaxiques, locutions ou expressions figées qui infléchissent, voire modifient totalement leurs significations) sous l'angle de leurs répétitions éventuelles dans le corpus.

Séquences d'unités

Au plan lexical, par exemple, les récurrences d'unités comme : *sécurité sociale, niveau de vie,* etc., sont dotées, dans les textes socio-politiques, d'un sens que l'on ne peut déduire à partir du sens des formes qui les composent. On appelle *segment répété* toute suite d'unités textuelles reproduite sans variation à plusieurs endroits d'un corpus. Le nombre des unités qui composent le segment est sa *longueur*.

On peut recenser les segments répétés constitués par les unités qui relèvent de chacun des systèmes d'annotation dont on dispose sur le texte. Les suites de catégories grammaticales, par exemple, considérées sous l'angle de leur répétition dans le corpus renseignent sur la fréquence relative des constructions syntaxiques[7].

La recherche systématique des segments répétés de ***Mitterrand1***, parmi les formes lexicales, fait ainsi apparaître un très grand nombre de récurrences de fréquence élevée. Tous ces constats de répétition ne renvoient pas au même niveau d'analyse linguistique. Certains résultent de l'utilisation de syntagmes relativement bien formés, d'autres sont produits par la reprise partielle dans des phrases différentes de fragments plus ou moins autonomes au plan syntaxique.

Dans le tableau 3, on a rassemblé quelques-uns des segments qui sont à la fois longs et fréquents dans ce corpus. La colonne L donne la longueur du segment mesurée en formes graphiques, la colonne F indique sa fréquence.

Tableau 3. — Quelques segments fréquemment répétés dans ***Mitterrand1***

[7] On s'étonne par exemple, lors de l'analyse d'***Enfants,*** de ne pas trouver de segments répétés comprenant des verbes dans les réponses spécifiques (cf. *infra*) des plus diplômés

L	F	segment
7	13	j ai dit tout à l heure
7	11	l ai dit tout à l heure
6	42	il n y a pas de
6	15	ce n est pas moi qui
6	15	je suis président de la république
6	15	que le président de la république
5	106	il n y a pas
5	93	le président de la république
5	36	dit tout à l heure
5	36	mais ce n est pas
5	34	de ce point de vue
4	366	ce n est pas
4	211	président de la république
4	190	je n ai pas
4	146	il n y a
4	124	un certain nombre de
4	121	tout à l heure

Quasi-segments

A côté des séquences reprises à l'identique à plusieurs endroits du corpus, on trouve des séquences qui sont l'objet de reprises partielles : la séquence je {catégorie=verbe} fermement que, par exemple, peut se réaliser sous la forme je pense fermement que, je crois fermement que, etc. Bécue (1993) a proposé un algorithme qui repère des *quasi-segments* (répétés). Cet algorithme permet, par exemple, de rassembler en une même unité (faire {lemme=<1>}$^{+}$ sport) les séquences comme faire du sport et faire un peu de sport, etc. Cependant, les quasi-segments sont encore plus nombreux que les segments, et leur recensement pose des problèmes de sélection et d'édition.

Cooccurrences

Pour une unité-pôle donnée, plusieurs méthodes permettent de sélectionner d'autres unités textuelles qui ont fortement tendance à se trouver dans un même voisinage que cette unité[8]. Le principe général de ces méthodes est le suivant. Pour sélectionner les formes cooccurrentes d'une forme-pôle, on commence par définir une unité de contexte, ou voisinage, à l'intérieur duquel on considérera que deux unités sont cooccurrentes. Cette unité de contexte peut correspondre à la phrase ou encore être constituée par un contexte de longueur fixe (*k* occurrences avant, et *k* occurrences après la forme-pôle). L'espace de cooccurrence peut également être défini de manière à ne pas dépasser les limites d'un constituant syntaxique. Si l'on se donne, à partir de l'exemple présenté plus haut (section 2.1), une fenêtre de deux occurrences avant et après la forme-pôle est (laquelle compte 2 occurrences), on construit autour de chacune des occurrences de la forme *est*, deux fenêtres matérialisées par les contextes compris entre les barres verticales :

```
Le 14              | juillet, c' est sans aucun |
```

[8] Les applications de ces méthodes à l'étude de cooccurrences entre d'autres unités linguistiques devront faire l'objet d'études au cas par cas.

```
doute
sans aucun doute | - et c' est fort important | -
l'
```

Dans ce cas, on sélectionne les cooccurrences de la forme-pôle avec les formes : juillet, c', sans, aucun, et , c', fort, important. Si l'on décide, toujours à partir de ce même extrait, de borner l'espace de cooccurrence au syntagme nominal minimal autour de la forme-pôle notre, on obtient une cooccurrence unique avec la forme armée.

Plusieurs méthodes statistiques se fixent pour but l'extraction des cooccurrences les plus remarquables dans un corpus de textes. Cette extraction s'appuie en général sur la comparaison des sous-ensembles de contextes qui contiennent l'unité-pôle avec ceux desquels elle est absente. Pour chaque unité-pôle, on sélectionne ainsi un ensemble d'unités qui se trouvent situées de manière privilégiée dans les mêmes unités de contexte[9].

Filtrage des résultats

La sélection automatisée des segments répétés, quasi-segments et cooccurrences fréquemment attestés dans un corpus produit des listes d'unités qui renvoient en général à des niveaux très différents de l'analyse linguistique (lexies plus ou moins figées, tournures syntaxiques récurrentes, tournures de rhétorique etc.). Pour réduire le volume des listes ainsi constituées, certains chercheurs ont entrepris de constituer des procédures de filtrages applicable à ces listes afin d'en extraire, par exemple, les seuls éléments qui correspondent à des syntagmes bien formés :

~~ce n est pas moi qui~~

je suis président de la république

~~que le président de la république~~

Comparer des décomptes au sein d'un corpus partitionné

Pour apprécier la répartition d'une unité linguistique à l'intérieur d'un corpus, il est nécessaire d'établir des comparaisons avec l'ensemble des unités de même type contenues dans le corpus. Une unité ne peut être jugée fréquente (ou rare) dans un texte que par comparaison avec d'autres unités dans ce même texte ou dans d'autres textes.

En pratique, ces comparaisons sont souvent malaisées du fait qu'il faut apprécier des décomptes qui concernent des unités dont les fréquences varient fortement dans des textes dont la longueur peut elle-même être très variable.

Le logiciel THIEF d'Étienne Brunet permet, par exemple, d'étudier la répartition de chacune des formes attestées dans le corpus du *Trésor de la Langue Française* parmi dix tranches chronologiques prédéfinies. On trouve figure 3 l'histogramme d'un indice qui permet de juger de la répartition de la forme *gloire*

---

[9] Lafon (1984) et Labbé (1990) proposent des méthodes destinées à extraire les couples d'unités lexicales qui se rencontrent souvent à l'intérieur d'une même phrase. Church et Hanks (1990) utilisent, dans le même but, l'information mutuelle issue de la théorie de la communication de R. Shannon.

dans ces dix tranches[10].

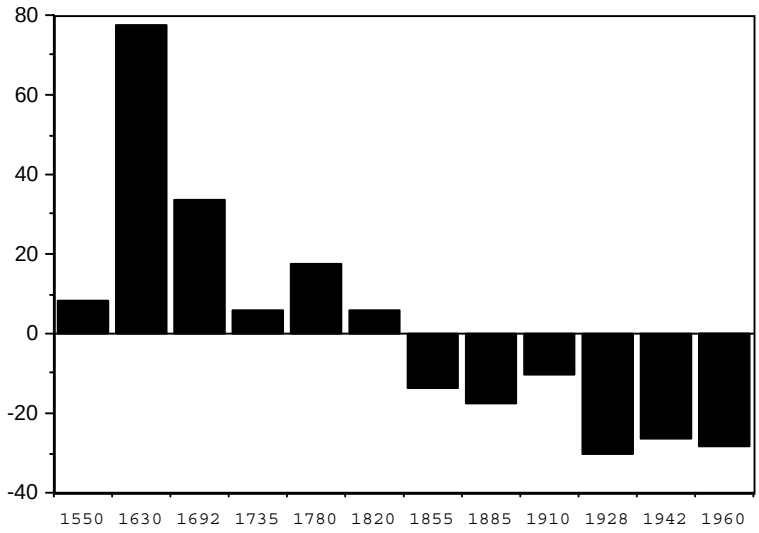


Figure 3. — La forme *gloire* dans dix tranches du TLF

Cette représentation graphique du phénomène appelle une interprétation très simple. La forme est tombée dans une désuétude relative au fil des périodes considérées.

La multiplication de résultats de ce type, à propos de formes différentes, incite à poser au corpus des questions plus générales. Quelles sont les formes qui subissent un sort similaire au cours des mêmes périodes ? Quelles sont celles qui au contraire voient le nombre de leurs occurrences augmenter relativement ?

Pour répondre de manière plus globale à des questions de ce type, il faut recourir aux méthodes de la statistique multidimensionnelle. Le point de départ des différentes méthodes qui servent à organiser la description comparative des parties d'un corpus est un tableau à double entrée que l'on constitue en croisant les parties du corpus et les différents types qui constituent le système d'unités préalablement choisi.

[10] Le calcul d'écart-réduit employé ici compare l'écart de la répartition observée dans chaque tranche à une répartition théorique.

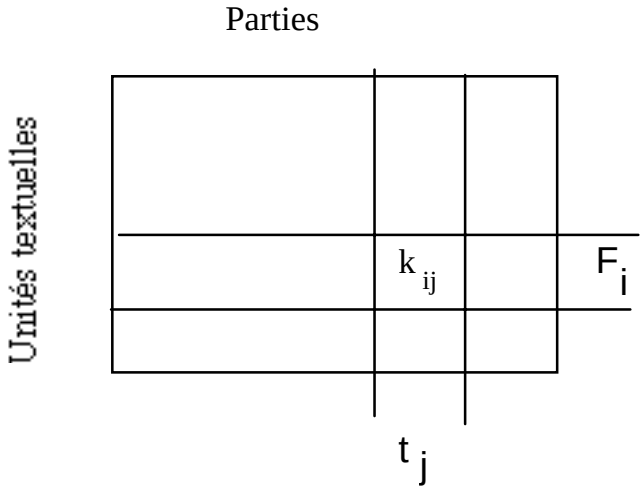


Figure 4. — Tableau de départ pour les analyses statistiques

A l'intersection de la ligne correspondant à l'unité *i* et de la colonne correspondant à la partie *j*, on trouve un nombre k_{ij} égal à la fréquence de l'unité *j* dans la partie *i* du corpus. La fréquence de l'unité *i* dans le corpus est égale à F_i. La longueur de la partie *j* (somme de toutes les occurrences de la partie j est égale à t_j.

Organiser la partition du corpus

A partir d'un même corpus, il est possible de constituer toute une série de partitions différentes (par émetteur ou par groupe d'émetteurs, si le corpus est plurilocuteur, en fonction de la date de rédaction, etc.). On peut ensuite décrire chacune des parties ainsi constituées par des systèmes de décomptes faisant intervenir des unités de différents niveaux (lemmes, formes graphiques, catégories grammaticales, ou tout autre type d'annotation). Le problème de la partition effective du corpus revêt une importance toute particulière dans la mesure où il s'agira ensuite d'étudier le contraste entre les parties découpées dans le corpus. La partition réalisée, on n'observera ensuite que des différences entre fragments du corpus ayant fait l'objet d'un même regroupement.

De son côté, la sélection d'un système d’unités linguistiques organise la comparaison des parties sur un plan d'analyse déterminé par les objectifs de la recherche. Les paragraphes qui suivent exposent brièvement les principes généraux du fonctionnement de ces méthodes sur des exemples empruntés à ***Enfants***.

En regroupant, par exemple, au sein d'une même partie les réponses fournies par les individus qui ont obtenu un diplôme équivalent, on réalise une partition du corpus en trois parties (Aucun, Baccalauréat, Supérieur). Cette partition permet ensuite d'étudier les variations entre agrégats de réponses.

Repérer les faits saillants

La méthode des spécificités (Lafon, 1980) permet de mettre en évidence les cases du tableau de départ dont l'effectif est particulièrement élevé (spécificités positives) ainsi que celles dont l'effectif est au contraire anormalement faible (spécificités négatives). Elle s'applique successivement à chacune des cases du tableau décrit plus haut. Pour calculer le diagnostic relatif à l'effectif constaté pour une unité dans une partie donnée, on prend en compte la comparaison de quatre

nombres :

k_{ij} – sous-fréquence de l'unité dans la partie considérée.

F_i – fréquence de l'unité dans l'ensemble du corpus.

t_j – nombre des unités dans la partie

T – nombre total des unités du corpus

Un calcul de type probabiliste permet de porter un jugement sur l'effectif contenu dans la case analysée (k_{ij}) compte tenu des trois autre nombres (F_i, t_j, T). Si l'effectif k_{ij} se situe dans les limites de ce que le calcul permettait d'espérer, on dit que la répartition constatée est *banale* (ce que l'on note « b »). Si ce n'est pas le cas, on calcule un indice de spécificité de la forme : +/-xx où :

\+ indique une spécificité positive (sur-représentation par rapport à ce que les nombres (F_i, t_j, T) laissaient prévoir ;

– indique une spécificité négative (sous-représentation) ;

xx est un indice de spécificité qui est d'autant plus élevé que la sous-fréquence analysée s'écarte d'une répartition « neutre » qui est sous-jacente au modèle des spécificités[11].

Les constats de spécificités établis pour une même unité à propos de chacune des parties du corpus permettent de décrire le comportement de cette unité au sein du corpus. On voit ci-dessous les diagnostics de spécificités obtenus dans chacune des parties pour la forme *problèmes* qui compte 108 occurrences dans l'ensemble du texte.

	Aucun	Baccalauréat	Supérieur	Total
problèmes	41	20	47	108
diagnostic	-03	b	+04	
effectif (= t_j)	8006	3111	4487	15604

Ces résultats indiquent que la forme graphique *problèmes* est sous-représentée (-03) chez les sujets sans diplôme. Elle est au contraire sur-représentée (+04) chez les plus diplômés. La notation b en regard de la catégorie Baccalauréat indique que l'effectif des occurrences de *problèmes* dans cette catégorie n'est ni excessivement élevé ni excessivement bas. Nous verrons plus loin comment organiser entre eux les différents constats de ce types obtenus à partir de différents systèmes d'unités.

Tableau 4. — Formes spécifiques pour les répondants les plus diplômés

[11] Le modèle probabiliste utilisé pour juger de cette répartition est ici le modèle hypergéométrique, couramment utilisé dans ce type d'application.

	F	f	Sp.
sur-emplois			
financières	174	79	+06
problèmes	108	47	+04
et	205	77	+03
face	10	8	+03
fait	25	14	+03
couple	95	39	+03
raisons	178	66	+03
affective	12	8	+03
difficultés	83	37	+03
responsabilités	22	13	+03
sous-emplois			
vie	180	35	-03
NON-REP	65	10	-03
le	474	111	-03
n	94	16	-03
vois	20	0	-03
manque	160	29	-03
aucune	33	3	-03
sais	25	1	-03
y	57	7	-03
faire	22	1	-03
pas	325	71	-03
emploi	79	13	-03
a	74	12	-03
travail	152	26	-04
il	105	15	-04
chômage	285	52	-05

Une fois ce calcul effectué pour chacune des cases du tableau analysé, le regroupement des diagnostics relatifs à une même partie fournit une description de cette partie par la mise en évidence des termes qu'elle sur-emploie, ainsi que celle des termes qu'elle sous-emploie[12]. Voici, à titre d'exemple, dans le tableau 4 ci-dessous, les formes jugées spécifiques, c'est-à-dire les formes tout particulièrement sur-représentées (resp. sous-représentées) dans la partie du corpus qui correspond aux plus diplômés.

Approches multidimensionnelles

Chacune des dimensions du tableau rectangulaire considéré plus haut permet de définir des distances (ou des proximités) entre les éléments de l'autre dimension[13]. Ainsi, l'ensemble des colonnes (dans notre cas les parties du corpus) permet de définir à l'aide de formules appropriées des distances entre lignes (ici les unités appartenant à un système d'annotation). De la même façon, l'ensemble des lignes permet de calculer des distances entre colonnes.

---

[12] On trouve un panorama des applications de ces méthodes aux textes socio-politiques dans (Habert, 1985).

[13] En analyse des données, on utilise souvent une distance qui est une somme de carrés pondérés dite *distance du chi-deux*. Cette distance possède toute une série de propriétés particulièrement intéressantes (Lebart et Salem, 1994, p. 87).

On obtient ainsi des tableaux de distances, auxquels sont associées des représentations géométriques complexes décrivant les similitudes existant entre les lignes et entre les colonnes des tableaux rectangulaires à analyser.

Le problème est alors de rendre assimilables et accessibles à l'intuition ces représentations, au prix d'une perte de l'information de base qui doit rester la plus petite possible.

Deux familles de méthodes permettent d'effectuer ces réductions :

- *Les méthodes factorielles* produisent des représentations graphiques sur lesquelles les proximités entre points-lignes et entre points-colonnes traduisent les associations statistiques entre lignes et entre colonnes ;
- *Les méthodes de classification* opèrent des regroupements en classes (ou en familles de classes hiérarchisées) des lignes ou des colonnes.

Classer les unités et les textes

Les méthodes de classification ascendante hiérarchique s'appliquent aux tableaux à double entrée décrits plus haut. On peut soumettre à la classification soit l'ensemble des colonnes du tableau (qui correspondent la plupart du temps aux différentes parties d'un corpus) soit celui des lignes de ce même tableau (lesquelles correspondent en général à un système d'unités textuelles recensées dans le corpus).

Classification ascendante hiérarchique

Dans le cas de la *classification ascendante hiérarchique*, on part d'un ensemble de *n* éléments, affectés chacun d'un poids proportionnel à leur importance dans l'ensemble, et entre lesquels on a calculé des distances. On commence par agréger les deux éléments les plus proches. Ce couple constitue alors un nouvel élément dont on peut recalculer à la fois le poids et les distances par rapport chacun des éléments qu'il reste à classer[14]. À l'issue de cette étape, le problème se trouve ramené à celui de la classification de *n-1* éléments. On agrège à nouveau les deux éléments les plus proches, et l'on réitère ce processus (*n-1* fois au total) jusqu'à épuisement de l'ensemble des éléments.

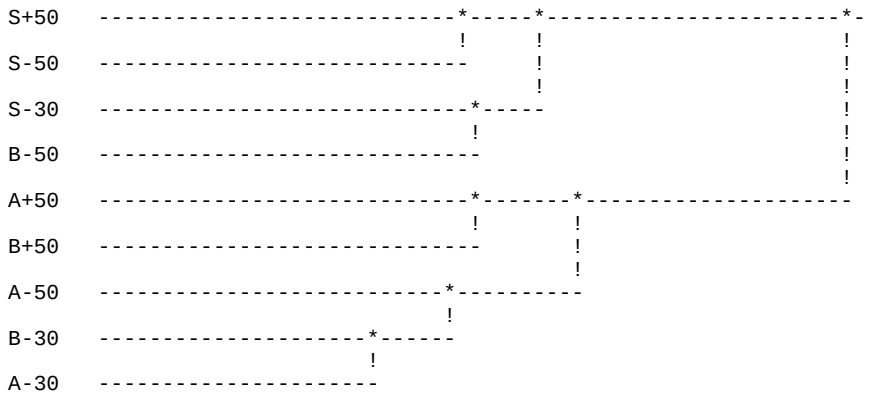
Chacun des regroupements effectués en suivant cette méthode s'appelle un *noeud*. L'ensemble des éléments terminaux rassemblés dans un noeud est une *classe*. La représentation de la classification sous forme d'arbre hiérarchique ou *dendrogramme* est la représentation la plus courante. L'interprétation d'une telle hiérarchie s'appuie sur l'analyse des seules distances entre éléments ou classes faisant l'objet d'un même noeud (*i.e.* seules les proximités entre éléments appartenant à une même classe peuvent être interprétées).

Appliquée au tableau analysé ci-dessus, la classification ascendante hiérarchique produit un regroupement en deux sous ensembles relativement distincts : les diplômés du supérieur d'une part et les sans-diplômes d'autre part. Les groupes de diplômes intermédiaires se répartissant entre ces deux sous-ensembles.

Tableau 5 — Classification sur les parties d'***Enfants***

---

[14]Dans la pratique il existe un grand nombre de façons de procéder qui correspondent à cette définition, ce qui explique la grande variété des méthodes de classification automatique, sur ces méthodes on peut consulter (Saporta, 1990, p. 241-261).



Les classifications effectuées sur l'ensemble des parties et celles réalisées à partir de l'ensemble des unités, répondent à des besoins d'analyse distincts qui entraînent, dans les deux cas, des utilisations différentes de la méthode.

Classifications de formes

Lorsqu'il s'agit d'étudier des textes (littéraires, politiques, historiques), les classifications portant sur les formes d'un corpus concernent en général des ensembles dont la dimension dépasse très largement celle de l'ensemble des parties. L'arbre de classification réalisé à partir d'un tel ensemble se présente sous une forme relativement volumineuse qui complique considérablement toute synthèse globale. Dans la pratique, on abordera l'étude des classifications ainsi réalisées en considérant par priorité les associations qui se réalisent aux deux extrémités du dendrogramme :

- les classes du niveau inférieur de la hiérarchie constituées par des agrégations de formes agrégées dès le début de la classification et qui correspondent souvent à des associations de type cooccurrentielles ;
- les classes supérieures, souvent constituées de nombreuses formes, que l'on étudiera globalement.

Les associations réalisées aux premiers niveaux de la classification regroupent, par construction, des ensembles de formes dont les profils de répartition sont très similaires (proportionnels et parfois mêmes identiques) dans les parties du corpus. Le retour systématique au contexte permet seul de distinguer parmi ces associations celles qui proviennent essentiellement de la reprise de segments plus ou moins longs, celles qui sont générées par les cooccurrences répétées de plusieurs formes à l'intérieur de mêmes phrases ou de mêmes paragraphes et les associations qui résultent de l'identité plus ou moins fortuite de la ventilation de certaines formes.

La figure 6 montre une petite partie de l'arbre de classification réalisé à partir des formes les plus fréquentes dans ***Enfants.*** L'analyse du contenu de ces classes se fait en retournant fréquemment au contexte.

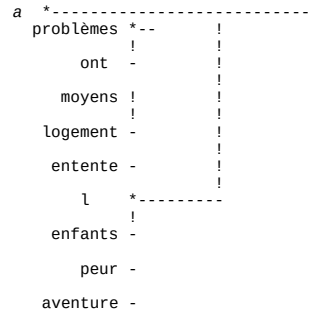


Figure 6. — Extrait d'une classification sur les formes d'***Enfants***

Classifications descendantes

Certains auteurs (Reinert, 1990) utilisent d'autres procédures de classification pour analyser les corpus textuels. Le principe général de la méthode est le suivant. On commence par découper dans le texte des *unités de contexte* (la plupart du temps, une fenêtre comportant quelques occurrences à gauche et à droite de chaque occurrence du texte). L'ensemble de ces unités est ensuite divisé successivement en classes (de manière dichotomique à chaque étape). Ce processus aboutit à rassembler des formes qui ont tendance à se retrouver dans des contextes proches.

L'approche factorielle

L'analyse factorielle des correspondances crée une typologie qui porte à la fois sur l'ensemble des parties du corpus et sur l'ensemble des unités par lequel ce dernier est décrit[15]. Négligeant toute une partie de l'information contenue dans le tableau des distances, cette méthode fournit des représentations approchées des distances calculées entre les éléments de chacun des deux ensembles mis en correspondance. Les graphiques-plans qui sont un des résultats fournis par l'analyse sont en quelque sorte les meilleures représentations bidimensionnelles possibles de chacun des ensembles. Sur ces graphiques, deux parties sont proches si elles emploient les mêmes unités dans des proportions semblables.

Cette méthode permet de créer une typologie qui peut s'affiner au fur et à mesure de la prise en compte des axes factoriels successifs. Elle est particulièrement adaptée à la mise en évidence des principales oppositions qui sous-tendent le corpus.

Remarquons que la classification ascendante hiérarchique et l'analyse factorielle sont des méthodes très complémentaires dans la mesure où l'une permet au chercheur de concentrer son attention sur les proximités locales pouvant exister entre chaque élément alors que la seconde rend compte des grandes oppositions pouvant exister dans le corpus.

Ainsi, les réponses contenues dans ***Enfants*** ont été regroupées cette fois en neuf parties qui correspondent au croisement de trois catégories de diplôme (A=aucun, B=Baccalauréat ou BEPC et S=Supérieur) avec trois catégories d'âge (moins de 30 ans, 30 à 50 ans, 50 ans et plus). On a ensuite calculé le tableau qui

---

[15] L'ouvrage de référence est le livre de J.-P. Benzécri et coll. (Benzécri, 1973). On trouvera des présentations différentes de cette même méthode destinées au lecteur non-mathématicien dans (Salem, 1987) ainsi que dans (Lebart et Salem, 1994).

croise ces neuf catégories avec les formes du corpus[16].

Commençons par un exemple très simple. On a représenté (Figure 7) les neuf parties du corpus en fonction de leur utilisation des formes : *raisons* (axe vertical) et *problèmes* (axe horizontal). La valeur portée sur chacun des axes est égale à la proportion d'utilisation (exprimée en 10 000èmes) de chacune de ces formes par chacune des parties. On voit que les parties ne se répartissent pas sur l'ensemble du graphique mais sont plutôt regroupées autour d'une des diagonales. Cela veut dire que l'emploi des deux formes par les émetteurs manifeste une *corrélation*. Ceux qui emploient beaucoup l'une des formes (S-30, S-50, c'est-à-dire les diplômés les plus jeunes) ont tendance à utiliser également l'autre (et inversement).

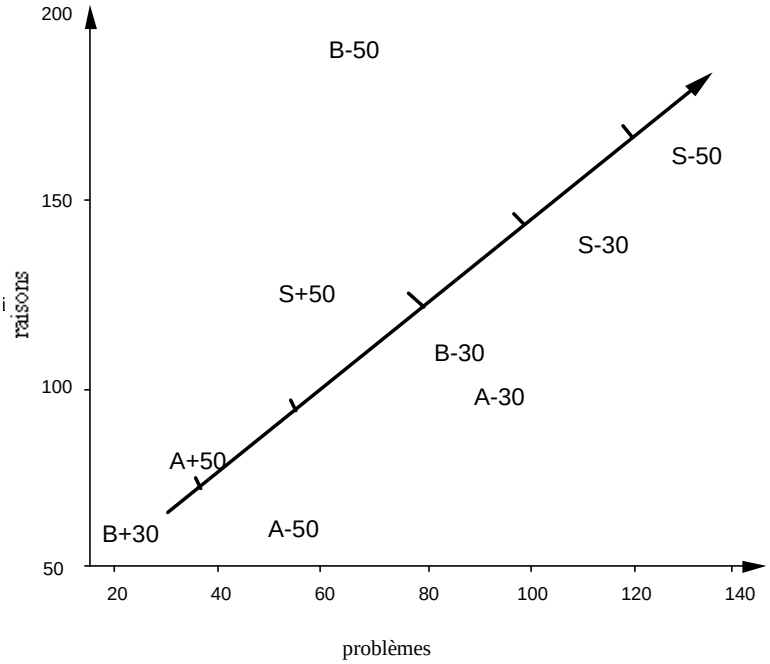


Figure 7. – Les parties d'*Enfants* et les formes *raisons* et *problèmes*.

Si l'on accepte de perdre un peu de l'information contenue sur ce graphique, on peut simplifier la représentation des parties en traçant un axe qui épouse *le mieux possible* la forme du nuage de points représenté sur la figure 7. Si l'on munit cet axe d'un système de coordonnées, on obtient une représentation des distances entre les parties (figure 8) qui est moins précise mais plus *synthétique*.

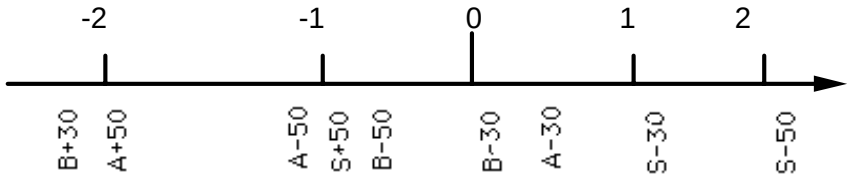


Figure 8. — Les mêmes parties disposées sur un « facteur »

Les méthodes factorielles opèrent, à partir des immenses tableaux soumis à

[16] Pour alléger les résultats, seules les formes de fréquence supérieure à 10 occurrences ont été retenues. L'expérience montre que ce type de sélection a peu d'influence sur les résultats de l'analyse.

l'analyse, des synthèses du même type. Partant d'un tableau qui compte cette fois plusieurs milliers de formes et toujours neuf parties, l'analyse des correspondances extrait une information synthétique. La représentation simplifiée des distances entre catégories met en évidence la principale information contenue dans le tableau de données soumises à l'analyse : la proximité (basée sur un usage proche du stock des formes lexicales) des agrégats proches par le diplôme ou par l'âge (figure 9).

Il faut comprendre que la méthode de calcul ne s'appuie à aucun moment sur des données extérieures lui permettant d'inférer des proximités entre tel ou tel agrégat. Les rapprochements sont effectués uniquement à partir des comparaisons du stock de vocabulaire employé par les répondants appartenant à un même agrégat âge / diplôme.

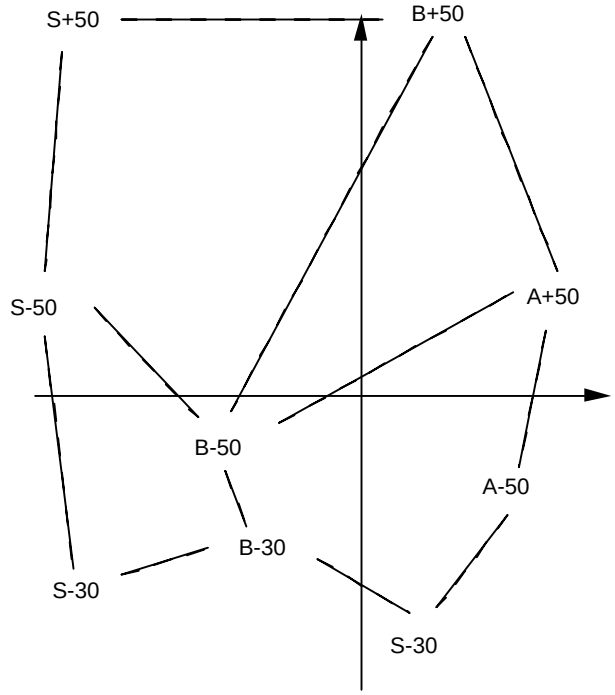


Figure 9. — Les 9 classes Age x Diplôme sur le plan des deux premiers facteurs de l'analyse.

Une représentation simultanée des formes et des parties sur le même graphique peut permettre de mettre en évidence les formes qui sont principalement responsables de cette typologie.

Articuler des constats sur des unités différentes

L'articulation des résultats obtenus à l'aide de telles méthodes à partir de différentes normes de dépouillement permet une description beaucoup plus sûre des contrastes entre les parties du corpus[17]. La typologie réalisée sur les parties dépend peu, dans le cas qui nous préoccupe, des variations dans la norme de dépouillement (lemme / formes graphiques, etc.). Loin de constituer une gêne

[17] Des résultats tout à fait similaires ont été obtenus dans une expérience du même type portant cette fois sur des décomptes de lemmes au sein de la même partition du corpus.

pour l'interprétation, les éclairages complémentaires projetés par différent systèmes d'unités nous aident à mieux comprendre les oppositions pouvant exister entre les textes que l'on compare.

Articuler unités isolées et séquences d'unités

L'exemple qui suit montre comment articuler de tels décomptes dans le cadre de la méthode des spécificités, la plus simple des méthodes exposées jusqu'ici.

Les occurrences du segment répétés problèmes financiers peuvent être considérées comme un sous-ensemble des occurrences de la forme problèmes pour lesquelles une occurrence de la forme financiers apparaît immédiatement après. On peut appliquer au segment répété problème financiers le calcul des spécificités.

Pour les deux formes et le segment évoqués, ce calcul donne :

Forme / diplôme	Aucun	BACC	Sup.	F
problèmes	41 -03	20 b	47 +04	108
financiers	37 b	19 b	30 b	86
problèmes financiers	17 -03	11 b	23 +03	51

Comme on le voit, les diagnostics ci-dessus ne coïncident pas tous entre eux. Ils rendent compte de la diversité des associations réalisées dans le corpus. La forme financiers, par exemple, est considérée comme régulièrement répartie alors que le segment problèmes financiers et la forme problèmes sont plutôt sur-représentés chez les plus diplômés.

Le tableau 6 interclasse d'après un indice de spécificité calculé selon les mêmes procédures des diagnostics obtenus sur des formes et sur des segments répétés dans le corpus. L'avantage de ce second tableau sur son homologue réalisé à partir des formes simples est qu'il constitue un pas, réalisé automatiquement, vers la remise en contexte des résultats.

Tableau 6. — Formes et segments les plus caractéristiques pour les répondants les plus diplômés

	F	f	Sp.
financières	174	79	+06
les difficultés financières	19	14	+05
difficultés financières	32	19	+04
problèmes	108	47	+04
fait de	10	7	+03
et	205	77	+03
face	10	8	+03
et les	17	10	+03
du couple	48	23	+03
fait	25	14	+03
situation économique	24	13	+03
raisons financières	93	38	+03
couple	95	39	+03
raisons	178	66	+03
problèmes financiers	51	23	+03
affective	12	8	+03
les problèmes	35	18	+03
difficultés	83	37	+03
des responsabilités	13	9	+03
responsabilités	22	13	+03
le fait	16	11	+03

Ce tableau présente de nombreuses redondances qui résultent du fait que, dans un premier temps, les listes d'unités spécifiques sont produites de manière entièrement automatique, sans aucun filtrage. L'illustration par les segments répétés précise la signification des unités mises en évidence par le calcul des spécificités. L'implication des dénombrements portant sur les segments répétés permet d'extraire de l'enchevêtrement inextricable des segments répétés des unités qui précisent la description par les unités effectuée à partir des unités isolées de leur contexte immédiat.

Articuler différents systèmes d'unités

La comparaison entre les différentes parties d'un corpus devient encore plus lisible si l'on implique les décomptes réalisées pour chacune d'elles à l'intérieur de différents systèmes d'unités linguistiques[18].

De la même manière que nous l'avons fait ci-dessus, il est possible de compléter la description des parties du corpus par des comptages réalisés sur l'ensemble des annotations disponibles dans le corpus considéré. Le tableau 7 montre les mêmes opérations de sélection d'unités caractéristiques réalisées cette fois à partir des annotations de type grammatical et des segments constitués à partir de ces dernières.

Tableau 7. — Formes graphiques, lemmes, catégories grammaticales et segments répétés les plus caractéristiques pour les répondants les plus diplômés

[18] Cf. (Salem, 1987 ; 1993) et (Habert et Salem, 1995)

	unités		F	f Ind.
C	{nom} {adjectif}	863	312	+07
F	financières	174	79	+06
L	*financier virgule*	123	59	+06
F	les difficultés financières	19	14	+05
C	{nom} {adjectif} {ponctuation}	32	20	+05
L	*le difficulté financier*	19	14	+05
F	problèmes	108	47	+04
F	difficultés financières	32	19	+04
C	{adjectif} {coord} {adjectif}	20	13	+04
C	{coord} {adjectif}	26	16	+04
C	{nom} {adjectif}{coord} {adjectif}	19	13	+04
C	{determinant ind} {nom} {adjectif}	36	20	+04
L	*difficulte financier virgule*	12	10	+04
L	*que ce*	26	17	+04
L	*difficulte financier*	32	19	+04
L	*financier*	374	136	+04
L	*probleme*	145	60	+04
F	problèmes financiers	51	23	+03
F	couple	95	39	+03
F	responsabilités	22	13	+03
F	raisons financières	93	38	+03
F	situation économique	24	13	+03
F	affective	12	8	+03
F	du couple	48	23	+03
F	et	205	77	+03
F	monde	16	10	+03
F	des responsabilités	13	9	+03
F	difficultés	83	37	+03
F	les problèmes	35	18	+03
F	et les	17	10	+03

Légende : La colonne de gauche indique la nature des unités et séquences d'unités prises en compte selon le code suivant : F – formes graphiques, L – lemmes, C – catégories grammaticales.

Comme plus haut, les unités sélectionnées dans ce tableau l'ont été en raison de leur abondance particulière dans la partie du corpus qui correspond aux plus diplômés. L'interclassement des unités selon l'indice de spécificité calculé de la même manière sur tous les types d'annotations et sur les segments réalisés à partir de ces dernières permet de classer l'ensemble des constats du plus surprenant au plus banal.

La redondance s'est encore accrue mais la description est devenue plus beaucoup plus riche, faisant intervenir de plusieurs niveaux de l'analyse linguistique.

Temps lexical

Certains corpus réunis par échantillonnage au cours du temps d'une même source textuelle présentent dès le départ une homogénéité remarquable : les textes réunis sont produits dans des conditions d'énonciation très proches, parfois par le même locuteur. Leur étalement dans le temps doit permettre de mettre en évidence ce

qui varie au cours du temps. Nous appelons ces corpus des *séries textuelles chronologiques*. ***Mitterrand1*** constitue, nous l'avons vu, un corpus de ce type.

Dans le cas des telles séries, les résultats factoriels font apparaître un schéma d'évolution chronologique qui rend compte de l'existence d'une évolution. Les apparitions, disparitions ou fluctuations des formes s'effectuent de manière suffisamment organisée, au regard du temps, pour que les périodes consécutives apparaissent plus proches dans l'emploi qu'elles font du vocabulaire que les périodes séparées par un intervalle de temps plus long.

La figure 10 montre des résultats issus d'une AFC portant sur les formes de fréquence supérieure ou égale à 5 occurrences dans ***Mitterrand1***. On le voit, les périodes consécutives sont plutôt proches les unes des autres. L'ensemble des points dessine une ligne incurvée en son centre.

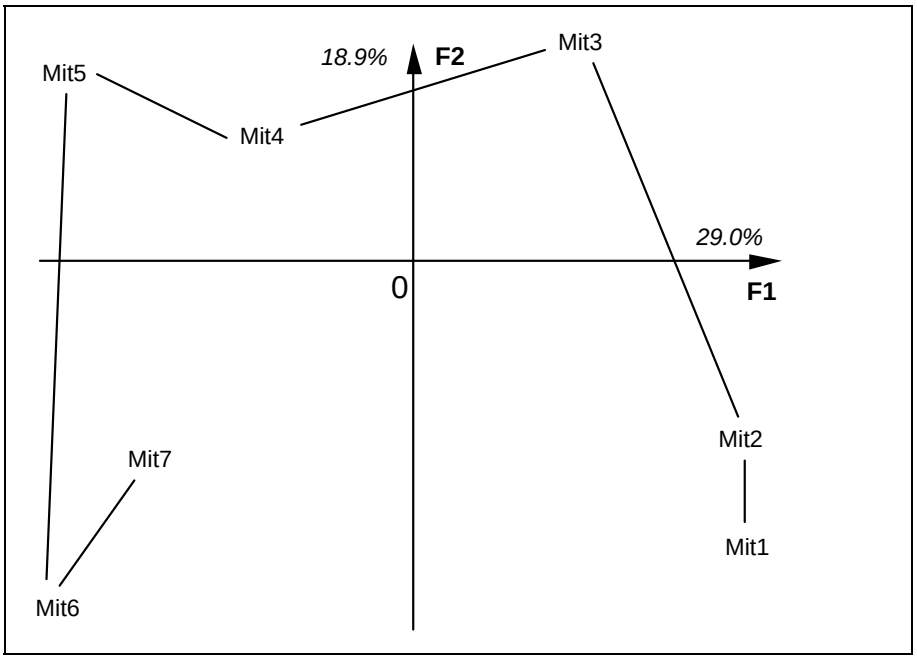


Figure 10. — Les deux premiers facteurs issus de l'analyse des correspondances[19]

Pour avancer dans l'analyse, il faut créer des procédures permettant d'exhiber les unités textuelles responsables de cette évolution d'ensemble.

Accroissements spécifiques

Le calcul des accroissements spécifiques permet de repérer les changements brusques dans l'utilisation d'un terme lors d'une période donnée par rapport à l'ensemble des périodes qui précèdent. Pour chaque terme dont la fréquence dépasse un seuil fixé à l'avance, pour chaque période du corpus à partir de la seconde, on compare, selon le modèle des spécificités présenté plus haut, la sous-fréquence observée dans la période considérée à la fréquence de cette même unité dans l'ensemble des périodes précédentes.

Le tableau 8 donne quelques accroissements spécifiques majeurs pour

[19] Il s'agit des deux premiers facteurs issus de l'analyse du tableau croisant formes graphiques de fréquence supérieure à 20 et périodes (1 397 formes x 7 périodes).

l'ensemble de ***Mitterrand1***. Les accroissements spécifiques sont notés à l'aide des symboles : / et \ qui indiquent des spécificités respectivement positive et négative de l'accroissement ; (*i.e.* un sur-emploi et un sous-emploi spécifique par rapport aux parties précédentes). La dernière colonne indique la période (pér.) *i.e.* la partie du corpus concernée par le diagnostic d'accroissement spécifique. Pour chaque terme, la colonne Fx donne le nombre des occurrences de ce terme dans le groupe de périodes précédentes.

Tableau 8. — Chronique des spécificités maximales pour ***Mitterrand1***

terme	F	Fx	f	spec.	pér.
nationalisations	42	31	0	/12	2
israel	71	56	2	\11	3
monsieur	430	213	91	/11	4
nouvelle calédonie	33	22	20	/11	4
référendum	27	19	18	/11	4
très	627	329	127	/11	4
chaîne	39	36	34	/19	5
la france	1016	722	106	\11	5
la majorité	91	70	45	/12	5
notre	442	337	35	\11	5
nous	2059	1700	308	\11	5
avons	523	488	30	\11	6
étudiants	28	28	27	/21	6
majorité	212	149	90	/20	5
nous	2059	1877	177	\17	6
oeuvres	29	24	19	/11	6
pour 100	204	195	2	\12	6
arabe	34	34	23	/13	7
l iran	50	50	41	/27	7
monde arabe	21	21	17	/12	7
nous	2059	2059	182	\12	7

Pour une période donnée, la liste des accroissements spécifiques de la période renseigne sur l'émergence d'un vocabulaire particulier. Le tableau 9 donne les accroissements ainsi calculés pour la 7e partie du corpus constituée par des interventions effectuées au cours des années 1987-1988.

Tableau 9. — Accroissements spécifiques majeurs pour la 7e période de ***Mitterrand1***

l iran	50	41	/27
iran	53	41	/25
arabe	34	23	/13
monde arabe	21	17	/12
d instruction	20	16	/11
instruction	23	17	/11
l irak	29	18	/09
irak	32	18	/08
élection	35	18	/07
président	303	73	/07
d armes	27	15	/07
un président	28	15	/07
politiques	105	34	/07
armes	93	32	/07
juge	35	17	/07
pays	748	151	/07
nous avons	413	27	\06
inflation	83	0	\06
avons	523	35	\07
jeunes	134	2	\07
nous	2059	182	\12

Formes chrono-homogènes

Les méthodes présentées ci-dessus permettent de décrire, au fil des périodes, l'évolution des unités textuelles que l'on peut recenser dans un corpus chronologique. Les schémas d'évolution établis pour chacune des unités font apparaître des ensembles d'unités qui ont tendance à évoluer de conserve au fil des périodes : les formes *chrono-homogènes*.

En fait, l'idée qui sous-tend cette approche est la suivante : pour des formes fréquentes dans le corpus, le fait que plusieurs formes évoluent de manière proportionnelle tout au long des périodes ne peut être mis au compte du hasard. Il faut donc, dans chaque cas, déterminer la cause profonde qui est à l'origine de ces regroupements. Selon les cas, on trouvera des groupements liés à une thématique, à une actualité, etc.

La figure 11 présente un groupe de formes, parmi les plus fréquentes de ***Mitterrand1***, qui sont chrono-homogènes par rapport à la forme je. On retrouve ici un ensemble de marqueurs de la première personne.

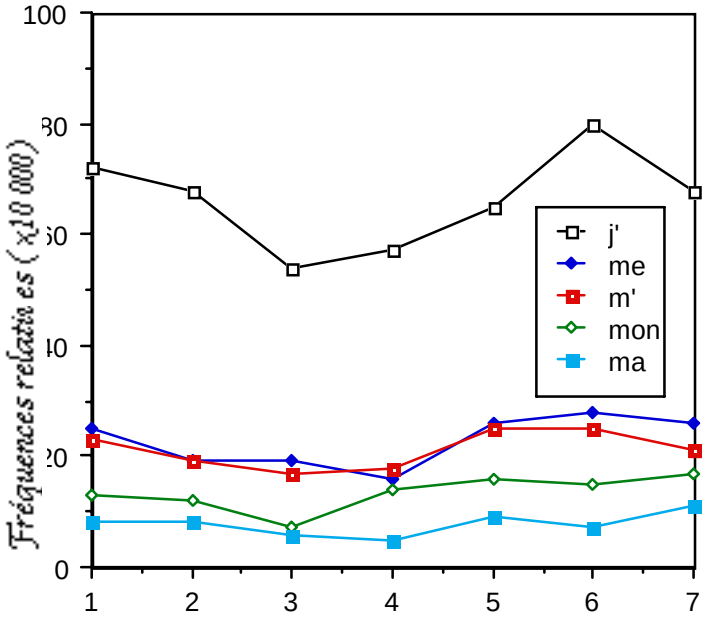


Figure 11. — Formes chrono-homogènes à la forme *je* dans ***Mitterrand1***

L'étude des séries textuelles chronologiques s'opère donc en combinant plusieurs types de méthodes. L'analyse des correspondances permet de vérifier que le corpus chronologique, compte tenu d'une périodisation donnée, relève bien du schéma général d'évolution du vocabulaire. Elle permet également de localiser des écarts éventuels avec le schéma général, qui seront dans la plupart des cas sources d'interrogations utiles. L'examen attentif des accroissements spécifiques signale à la fois des moments particuliers dans l'évolution du vocabulaire et les unités textuelles qui en sont à l'origine. Enfin, l'étude des termes chrono-homogènes permet de constituer des classes d'unités et d'étudier leur évolution conjointe au fil des périodes.

Conclusion

Les analyses portant sur des textes annotés apportent un complément d'information important, par rapport aux mêmes analyses effectuées à partir d'un découpage en formes graphiques, dès lors qu'il s'agit de mettre en évidence des unités textuelles caractéristiques pour chacune des parties d'un corpus de textes, encore que ces résultats soient difficiles à manier simultanément.

L'utilisation de comptages portant sur les segments répétés d'un corpus pour illustrer les typologies réalisées à partir des formes permet de dépasser les résultats obtenus sur les formes isolées de leur contexte immédiat et d'accéder à la description d'associations remarquables par leur répartition. Les différentes méthodes de calcul des cooccurrences concourent également à ce but.

Par exemple, dans le domaine de l'étude des textes politiques, l'expérience a montré que le singulier et le pluriel de certains substantifs renvoient souvent à des oppositions profondes au plan de l'idéologie politique. On peut dire que de grandes oppositions idéologiques se sont souvent exprimées à travers l'emploi du

singulier ou du pluriel d'une même forme de vocabulaire. *Les classes ouvrières*, proclamait le pouvoir monarchique sous Louis-Philippe (1830-1848) ; *la classe ouvrière,* contestaient les organisations ouvrières. De même les années 1970 ont vu s'opposer les défenseurs *des libertés républicaines* (la gauche et les syndicats) aux défenseurs de *la liberté* avec, bien entendu, des contenus partiellement différents. Cette distinction est en revanche moins pertinente dans le cas de l'étude de ***Menelas*** : le comportement du singulier et du pluriel de *sténose* ne justifie pas qu'on les considère séparément.

L'éclairage qu'apporte l'approche quantitative à la connaissance d'un corpus de textes réunis à des fins de comparaison s'exprime de manière privilégiée sous forme de contrastes entre les unités que l'on peut décompter dans les parties du corpus. Ces circonstances fournissent indirectement un critère quant au choix des unités à retenir dans les analyses textuelles : si les différentes réalisations d'une unité linguistique sont distribuées de la même manière parmi les parties du corpus que l'on compare, il ne sert à rien de les distinguer dans les comptages, car elles ne seront pas à l'origine des contrastes mis en lumière par les analyses statistiques. Si par contre les réalisattions d'une même unité ont des ventilations très différentes à l'intérieur du corpus considéré, le fait de les réunir en une même unité statistique prive le chercheur de constats qui auraient pu l'intéresser.