Faut-il prendre en compte la composition grammaticale des textes dans le calcul des spécificités lexicales ? Tests logométriques appliqués au discours présidentiel sous la V^{ème} République

Damon Mayaffre

CNRS – UMR, Bases, Corpus et Langage (Nice)

Abstract

Do we need to weight lexical specificities by the grammatical composition of texts ? Logometrical tests applied to the presidential discours under the 5th republic

For several years, textual statistics successively applied to untreated texts and then to the same lemmatized grammaticalized texts, has given global or identically synthetic results. Great textual classifications leading to more or less innovative textual typologies appear insensible to the linguistic unit processed (form or lemma, grammatical category or syntactic combination). Thus on an AFC for example, a given discourse standing out from others from a lexical viewpoint, will also stand out grammatically. Moreover, in this logic, the study of lexical specificities and then the study of grammatical specificities appear redundant more often than not, it is as if the same thing was measured twice.

What is at stake in this contribution is to manage to establish a relation between two intrinsically related linguistic events (the lexical event and the grammatical one) and to cast light - or in statistics terms to weight- one with the other.

So far the statistic probability to see such and such word appear has always been calculated in relation to the total set of forms, even if it is informed not by the total surface of the corpus and its sub-parts, but by the proportion of the grammatical category corresponding to that of the word in the corpus and its sub-parts.

Résumé

Depuis plusieurs années, la statistique textuelle appliquée successivement à des corpus de textes bruts puis aux mêmes textes lemmatisés/grammaticalisés donne des résultats globaux ou synthétiques identiques. Les grandes classifications de textes, qui aboutissent à des typologies textuelles plus ou moins innovantes, apparaissent insensibles à l'unité linguistique traitée (forme ou lemme, catégorie grammaticale et enchaînement syntaxique). Ainsi sur une AFC par exemple, un discours qui se distingue des autres d'un point de vue lexical s'en distinguera de manière similaire d'un point de vue grammatical.

L'enjeu de cette contribution est de réussir à mettre en relation deux événements linguistiques intimement liés (l'événement lexical et l'événement grammatical) et d'éclairer –c'est-à-dire en terme statistique, peut-être, de pondérer– l'un par l'autre.

Jusqu'ici, la probabilité statistique de voir apparaître tel mot a toujours été calculée par rapport à l'ensemble des mots du corpus, quand bien même cette probabilité est informée, non par la surface totale du corpus, mais par la proportion de la catégorie grammaticale dont les mots relèvent.

Mots-clés : statistique textuelle, spécificités, lemmatisation, logométrie, discours politique.

1. Introduction

Cette contribution naît d'une double réalité maintes fois constatée dans les travaux logométriques (i.e. traitements statistiques appliqués aux textes dans toutes leurs dimensions, graphique, lexicale, grammaticale et syntaxique).

1.1. Depuis plusieurs années, la statistique textuelle appliquée successivement à des corpus de textes bruts puis aux mêmes textes lemmatisés/grammaticalisés donne des résultats synthétiques identiques. Les travaux logométriques de Brunet (2000) ou Kastberg (2006) sur des corpus littéraires, ou les nôtres (Mayaffre 2004) sur des corpus politiques montrent que les grandes classifications des textes qui aboutissent à des typologies plus ou moins innovantes sont insensibles à l'unité linguistique traitée (forme ou lemme, catégorie grammaticale et enchaînement syntaxique). Sans présumer encore du facteur déterminant, on peut affirmer notamment qu'une partie du corpus (un locuteur, une œuvre, une année dans une séquence chronologique) qui se distingue lexicalement des autres parties du corpus, sur une analyse arborée, un tableau de distances lexicales, une analyse factorielle, s'en distinguera de manière similaire grammaticalement. Et vice versa. Ainsi, par exemple, les AFC du corpus présidentiel (1958-2002) calculées selon l'ensemble des mots et selon l'ensemble des codes grammaticaux donnent globalement des sorties machines ressemblantes lorsqu'on croise les deux premiers axes (figure 1 et 2).

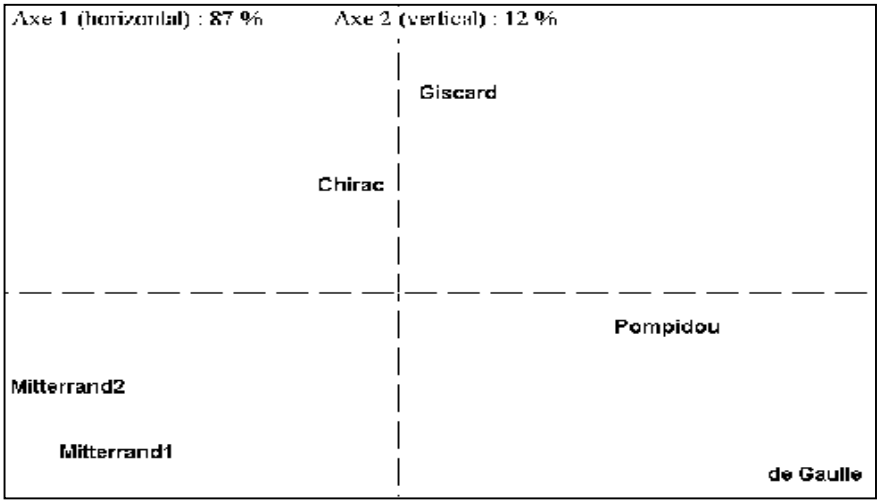

Figure 1 : AFC du corpus présidentiel (1958-2002) selon les mots utilisés

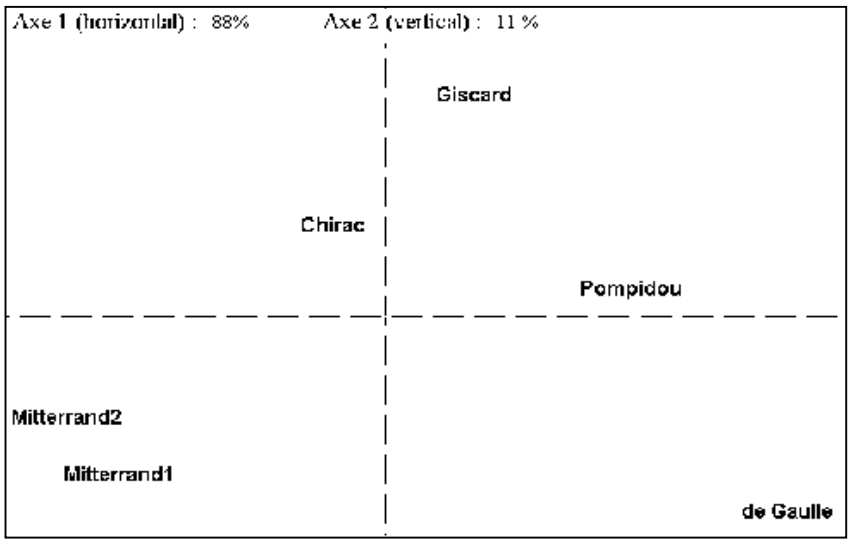

Figure 2 : AFC du corpus présidentiel (1958-2002) selon les codes grammaticaux utilisés

1.2. Concomitamment et de manière évidemment liée, l'étude des spécificités lexicales puis l'étude des spécificités grammaticales (ou encore l'étude des spécificités des combinaisons syntaxiques) apparaissent très souvent redondantes. Ainsi un chercheur, initié au genre du corpus qu'il traite, pourra prédire à la vue des 10 ou 20 premières spécificités lexicales d'un locuteur, ses spécificités grammaticales, et à la liste des spécificités grammaticales, deviner *grosso modo* une partie de sa liste de mots spécifiques. Dans le cadre du discours politique français du XX^{ème} siècle, un orateur politique de dimension nationale (président de la République, premier ministre, député, chef de parti), qui tient un *discours nominal* au sens de (Brunet, 1981) (sur-utilisation des noms, des déterminants et des adjectifs), sur-emploiera probablement les lexies "France" et "pays" (les deux noms les plus utilisés dans la langue politique française) ; de même, il sur-utilisera "de" et "la" (les déterminants les plus utilisés dans la langue en général). Inversement, un orateur qui sur-utilisera "France" ou "pays" sur-emploiera sans doute la catégorie nominale.

Ces exemples paraîtront naïfs. Ils pourraient pourtant être systématisés : un locuteur politique qui sur-utilise la catégorie verbale sur-utilisera *quasi* nécessairement les lemmes "être", "avoir" ou "dire", à moins que cela soit en utilisant "être", "avoir" ou "dire" qu'il sur-utilisera les verbes. Un auteur qui sur-utilise "je" sur-emploiera *quasi* automatiquement les pronoms, tant la classe des pronoms est faible et que la sur-utilisation d'un d'entre-eux suffit souvent, mathématiquement, pour sur-représenter la catégorie pronominale entière. Etc.

2. Hypothèses

2.1. D'apparence bénigne, la question est épineuse d'abord d'un point de vue linguistique. Elle nous paraît même insoluble sans un détour par la théorie linguistique. Nous nous proposons simplement ici d'évoquer les termes du problème en reportant la réflexion linguistique à une autre manifestation que les JADT.

Dans la bouche d'un locuteur, est-ce le choix grammatical qui informe le choix lexical ? Ou bien l'inverse ?

Sont-ce les pesanteurs grammaticales qui contraignent la liberté lexicale d'un locuteur ? Ou au contraire, le choix lexical qui décide la tonalité grammaticale du discours ?

La sur-présence statistique constatée de "la" dans un corpus doit-elle être considérée par l'analyste comme la *cause* ou la *conséquence* d'un discours à tonalité nominale ? Dit-on "je" *pour* pronominaliser son discours ? Ou sommes nous condamnés à dire "je" *parce que* l'on entend tenir un discours pronominal au procès d'énonciation tendu.

Une réponse instinctive à ces questions est que la contrainte grammaticale pèse sur le choix lexical comme la structure sur l'évènement et le système sur son actualisation. Mais il y a là une thèse à démontrer et sans doute à nuancer. Un homme politique n'a-t-il pas le sentiment, tout au contraire, de construire son discours autour de choix lexicaux majeurs ("peuple" ou "république", "nous" ou "vouloir faire", etc.) qui organisent le discours ?

2.2. D'un point de vue statistique, la question est tout aussi difficile. L'enjeu est de réussir à mettre en relation deux événements linguistiques intimement liés (l'événement lexical donc et l'événement grammatical) et d'éclairer –est-ce à dire, en termes statistiques, de "pondérer ? – l'un par l'autre.

Le but est double. D'abord, nous l'avons dit, il s'agit de ne plus témoigner deux fois de la même chose. La réalité linguistique semble former un tout atomique et il apparaît peu économe voire artefactuel de vouloir en témoigner différemment comme s'il était composé de

plusieurs réalités indépendantes les unes aux autres. Ensuite, et surtout, il s'agit d'essayer de percevoir les choses plus finement que ne le propose la lexicométrie traditionnelle.

L'hypothèse principale qui est donc testée est que la composition grammaticale écrase, d'un point de vue statistique, des réalités lexicales plus sensibles (car plus nombreuses).

Certains événements lexicaux qui ont résisté à cet écrasement en témoignent. Ainsi dans le discours présidentiel (1958-2002), le discours de Giscard (comme celui de De Gaulle et Pompidou) se caractérise par un sous-emploi important des verbes (figure 3):

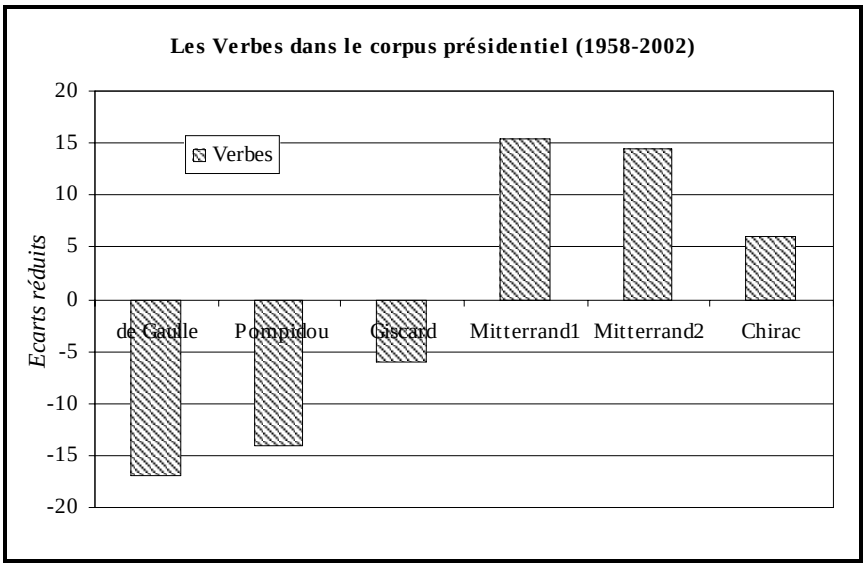

Figure 3 : Répartition des verbes dans le corpus présidentiel (1958-2002)

Plus généralement, la composition grammaticale du discours de Giscard (composition relative aux autres présidents, comme on le sait) se présente ainsi (figure 4) :

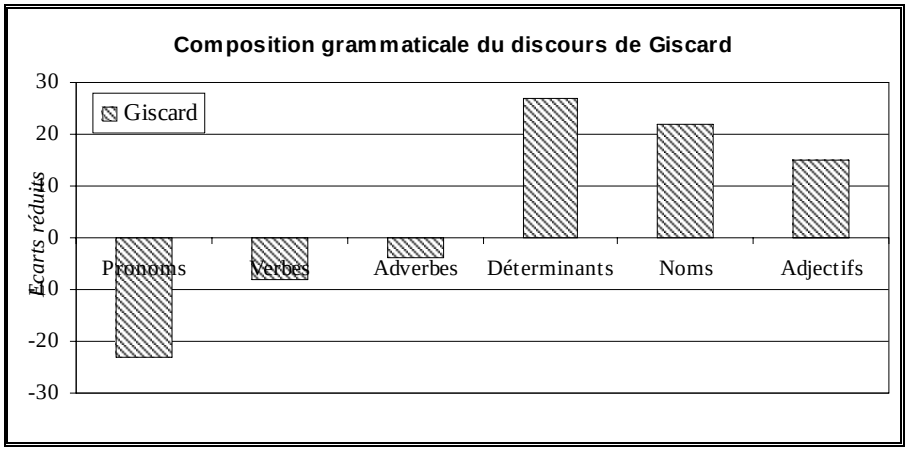

Figure 4 : Composition grammaticale du discours de Giscard

Ainsi – cause ou conséquence ? – dans la liste des spécificités lexicales positives giscardiennes, nous ne trouvons pratiquement que des noms, des déterminants et des adjectifs. Les 10 plus fortes spécificités en témoignent (tableau 1).

Lemmes	Occ. dans le corpus	Occ chez Giscard	Écarts réduits
1 - Actuel (adj.)	1247	684	29,7
2 - Situation (nom)	1663	749	24,3
3 - Heure (nom)	1643	686	21,0
4 – Problème (nom)	2870	1041	20,2
5 - Énergie (nom)	474	284	20,0
6 - Événement (nom)	97	96	18,3
7 - Un (déterminant)	44862	11036	18,2
8 - De (déterminant)	95156	22347	17,9
9 - Indiquer (verbe)	**351**	**208**	**16,6**
10 – Question (nom)	1866	689	16,5

Tableau 1 : Les 10 premières spécificités lexicales de Giscard dans le corpus présidentiel 1958-2002

Pourtant dans cet océan nominal, sur-nagent quelques spécificités verbales. Le verbe "indiquer" (en gras dans le tableau) se trouve même nettement sur-utilisé (écart réduit de +16,6) par Giscard en se situant dans le palmarès des spécificités lexicales du futur académicien. (figure 5).

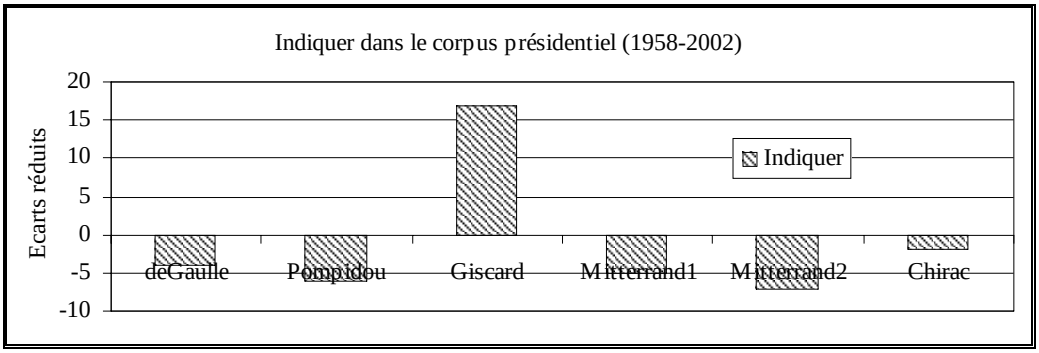

Figure 5 : Distribution de "indiquer" dans le corpus présidentiel (1958-2002)

La superposition intellectuelle de ces différents constats (figures 3 et 5 notamment) suggère la dimension extraordinaire du verbe "indiquer" ; situation dont la statistique ne nous dit rien faute de réussir à croiser lesdits constats. Giscard se prive (relativement) de verbe *et pourtant* sur-utilise "indiquer" dans de fortes proportions ; Giscard utilise moins de verbes que Chirac et Mitterrand *et pourtant* plus d'"indiquer" que ses condisciples (en valeur absolue, sur des corpus sensiblement identiques, entre 4 et 10 fois plus).

Quand bien même le déficit de verbes chez Giscard est en proportion bien plus faible que le laisse penser en terme de probabilité la figure 3 (14,53 % de verbes chez Giscard-le-nominal contre 15,9 % chez Mitterrand-le-verbal), *ce déficit grammatical structurel pèse, à notre avis, sur l'espérance statistique de voir l'apparition de tel ou tel verbe (ici "indiquer") dans le discours, et doit être pris en compte dans la caractérisation des spécificités lexicales.*

De manière plus générale, il est possible d'imaginer (et de calculer) un seuil critique où une forme ramenée à l'ensemble des formes se trouve spécifique alors qu'elle ne le serait pas *ramenée à*, ou *pondérée par*, la proportion de sa catégorie grammaticale. Inversement, certains mots ne figurant pas dans la liste des spécificités lexicales pourraient être pourtant considérés comme caractéristiques d'un locuteur au regard de l'utilisation que celui-ci fait de la catégorie grammaticale à laquelle appartiennent ces mots.

3. Propositions

L'objectif de cette contribution est donc d'explorer les moyens de prendre en compte la tonalité grammaticale des discours dans l'appréhension de leur composition lexicale.

3.1. Monière, Labbé et Labbé (2005) se sont confrontés récemment à cette question de "l'incidence des variations de densité des catégories grammaticales" (*ibid* : 94) sur la fréquence d'apparition d'un vocable et ont proposé un modèle dans la dernière livraison de *Corpus.* Leurs premiers résultats nous apparaissent si convaincants (*ibid* : 93-94) qu'ils nous encouragent dans la marche à suivre quand bien même nous proposons un autre modèle.

3.2. Jusqu'ici la probabilité statistique de voir apparaître *k* fois tel mot (ou tel lemme) a donc toujours été calculée par rapport à l'ensemble des formes (ou des lemmes) du corpus. Ainsi, selon le modèle hypergéométrique (Lafon, 1984), l'algorithme le plus communément accepté, est :

$$\text{prob}\,(x=k) = \frac{f!\,(T-f)!\;t!\,(T-t)!}{k!\,(f-k)!\,(t-k)!\,(T-f-t+k)!\;T!}$$

Où,

T = taille du corpus,

t = taille du texte,

f = fréquence du mot dans le corpus,

k = fréquence du mot dans le texte,

De la même manière, le calcul des spécificités, selon la loi normale, applique la formule rappelée par (Brunet, 2001) :

$$z = \frac{k - fp}{\sqrt[2]{fpq}}$$

Où,

k = fréquence du mot dans le texte,

f = fréquence du mot dans le corpus,

p = étendue du texte (t) / étendue du corpus (T),

q = 1 – p.

La solution alternative la plus radicale est donc de calculer désormais cette probabilité non par rapport à l'ensemble du corpus mais par rapport à l'ensemble des catégories. C'est-à-dire, où :

T(c) = taille de la catégorie considérée dans le corpus (T(v) = verbes, T(n) = noms, etc.),

t (c)= taille de la catégorie considérée dans le texte (t(v) = verbes, t(n) = noms, etc.)

Pour ce qui nous concerne, la matrice de nos données passe ainsi du tableau 2 au tableau 3.

	De Gaulle	Pompidou	Giscard	Mitterrand1	Mitterrand2	Chirac	**Total T**
Forme1	k(f1, Gaul)	k(f1, Pomp)	k(f1, Gis)	k(f1, Mit1)	k(f1, Mit2)	k(f1,Chir)	K(f1)
Forme 2	k(f2,Gaul)	…	…	…	…	…	…
…	…	…	…	…	…	…	…
"indiquer"	23	13	208	36	16	55	351
…	…	…	…	…	…	…	…
Forme *n*	k(f*n*, Gaul)	…	…	…	…	…	K(f*n*)
Total t	**224119**	**237536**	**410855**	**370182**	**340364**	**340926**	**1923982**

Tableau 2 : Matrice des données pour le calcul des spécificités traditionnelles

	De Gaulle	Pompidou	Giscard	Mitterrand1	Mitterrand2	Chirac	**Total T(v)**
Verbe1	k(v1, Gaul)	k(v1, Pomp)	k(v1, Gis)	k(v1, Mit1)	k(v1, Mit2)	k(v1,Chir)	K(v1)
Verbe 2	k(v2, Gaul)	…	…	…	…	…	…
…		…	…	…	…	…	…
"indiquer"	23	13	208	36	16	55	351
…	…	…	…	…	…	…	…
Verbe *n*	k(v*n*, Gaul)	…	…	…	…	…	K(v*n*)
Total t(v)	**30361**	**32740**	**59736**	**58247**	**53462**	**51928**	**286.474**

Tableau 3 : Matrice des données pour le calcul des spécificités grammaticalisées

Nous insistons sur le fait que cette solution expérimentale est radicale car ici l'influence de la catégorie grammaticale n'est pas seulement prise en compte mais totalement neutralisée : le poids des items lexicaux n'est mesuré qu'à l'intérieur des catégories grammaticales.

4. Résultats

Les résultats obtenus sur le gros corpus des discours présidentiels (1958-2002) sont étonnants et apparaissent idéaux dans la mesure où ils n'invalident pas nos pratiques habituelles, mais les tempèrent. Autrement dit, ils n'abolissent pas le calcul traditionnel des spécificités, clef de voûte de la plupart des études lexicométriques depuis 30 ans, mais le redressent.

Les listes comparatives des spécificités de Giscard calculées de manière traditionnelle sur l'ensemble des formes, puis en tenant compte des catégories grammaticales, sont, par exemple, très parlantes.

De manière modérée mais systématique, les représentants lexicaux des catégories grammaticales défavorisées chez Giscard (verbes, pronoms, adverbes) remontent dans la liste, lorsque les représentants lexicaux des catégories favorisées (noms, adjectifs, déterminants) descendent.

Le verbe "indiquer" par exemple situé jusqu'ici au 9ème rang des spécificités lexicales se trouve classé désormais 7ème (+2). De manière générale, les 40 premières spécificités lexicales de Giscard, calculées traditionnellement, comptent seulement 2 verbes ("indiquer" et "rechercher"), alors qu'elles en comptent 4 ("indiquer", "rechercher", "conduire" et "être") calculées en prenant en compte les catégories grammaticales. Sur les 70 premières spécificités lexicales, nous trouvons 6 verbes avec le premier mode de calcul contre 12 avec le second. Ainsi, apparaissent plus clairement des événements lexicaux (ici de nature verbale) qui nous semblent caractériser fortement le discours giscardien et qui risquaient de passer jusqu'ici inaperçus. "Considérer" (désormais 61ème rang, remontée de +14), "observer" (66ème rang, +15) ou "concerner" (69ème rang, +14)) par exemple, illustrent parfaitement le discours didactique du pédagogue Giscard qui se comporte plus en professeur qu'en président (Mayaffre, 2004 : 70-76). Plus loin encore, le verbe "poser" par exemple ("il faut se poser la question…", "je voudrais poser en premier point…") apparaît désormais comme spécifique de Giscard à la 138ème place et un écart réduit de +6,7 alors qu'il ne franchissait pas le seuil de significativité précédemment.

Grâce à une implémentation des nouveaux modes de calcul dans HYPERBASE, réalisée par Étienne Brunet cet hiver, le test a été généralisé à l'ensemble du corpus présidentiel, selon les deux modèles statistiques énoncés plus haut (loi hypergéométrique, loi normale) et aussi bien pour les spécificités positives que négatives. Il donne des résultats satisfaisants.

Chez les locuteurs verbaux, Mitterrand et Chirac, les spécificités lexicales verbales sont légèrement déclassées et les spécificités lexicales nominales revalorisés. Et inversement, chez les locuteurs nominaux (de Gaulle, Pompidou, Giscard).

Cela nous permet de voir par exemple que la sur-utilisation de "naturellement" par Chirac –1er mot spécifique du président même lorsque le calcul prend en compte la composition grammaticale du discours– relève bien d'un choix lexical et non d'une tendance générale, par ailleurs constatée, de Chirac à utiliser beaucoup d'adverbes. En revanche, l'usage que Chirac fait de "parfaitement" par exemple n'apparaît plus spécifique au regard du discours à tonalité adverbiale du président.

Au final, pour des raisons mathématiques à explorer, les seules anomalies concernent les catégories grammaticales qui comptent peu d'entrées (les déterminants et les pronoms) et/ou les lemmes qui comptent énormément d'effectifs ("le", "un", "être"). Dans ce cadre là, le redressement qu'opère le calcul des spécificités selon la catégorie grammaticale nous semble trop important et demande à être affiné. Chez Giscard, "être" dont les effectifs sont pléthoriques (100 fois plus que n'importe quel autre verbe) remonte trop spectaculairement passant de la 198ème place avec un écart réduit de 5,9 à la 47ème place avec un écart réduit de 10,4). De manière symétrique "un" se trouve trop fortement déclassé passant de la 7ème place (+18,3) à la 84ème place (+8,5).

5. Conclusion

La lemmatisation des textes n'est plus aujourd'hui ni un moyen coûteux (c'est rapide) ni un moyen dangereux (c'est fiable). La seule condition à sa pratique est que l'on continue d'accéder, toujours, au texte brut et de favoriser, d'abord, le traitement des formes graphiques : il s'agit-là du postulat de la logométrie (Mayaffre, 2005) et de la mise en œuvre que propose HYPERBASE en offrant systématiquement à l'utilisateur la possibilité de naviguer entre formes et lemmes, textes bruts et textes étiquetés.

Mais la lemmatisation nous paraît sous-exploitée si l'on n'en tire pas tous les bénéfices. Juxtaposer une étude des formes graphiques, une étude des formes lemmatisées, une étude des codes grammaticaux aboutit actuellement le plus souvent à des redites pour caractériser un discours. Cela est toujours confortable de doubler ou tripler les résultats mais il existe une forme de trompe l'œil à prétendre confirmer des constats lorsqu'il s'agit, en fait, seulement de les répéter.

La lemmatisation et les études logométriques prouveront tout leur intérêt seulement si elles permettent de croiser les résultats qu'elles obtiennent aux différents niveaux de descriptions linguistiques qu'elles proposent, c'est-à-dire en mettant en résonance les constats produits sur les lexies, les lemmes, les catégories grammaticales et les enchaînements syntaxiques.

Ce programme général nous paraît devoir être appliqué au plus célèbre calcul de la statistique textuelle, celui des spécificités.

Références

Brunet É (1981), *Le vocabulaire français de 1789 à nos jours d'après les données du "Trésor de la Langue Française"*. Genève-Paris, Slatkine-Champion, Vol. I, 852 p. ; vol. II, 518 p. ; vol. III, 453 p.

Brunet É. (2000). Qui lemmatise, dilemme attise, *Lexicometrica*, n°2.

Brunet É. (2001). *Manuel d'utilisation d'Hyperbase* (mise à jour mai 2001), Nice.

Brunet É. (2002). Le lemme comme on l'aime. In *JADT 2002*, Saint Malo, Irisa-Inria, : 221-233.

Habert B., Narazenko A. et Salem A. (1997). *Les linguistiques de corpus*. Paris, Colin.

Kastberg Sjöblom M. (2006). *L'écriture de J. M. G. Le Clézio – Des mots aux thèmes*. Paris, Champion.

Labbé C et Labbé D. (1997). Que mesure la spécificité du vocabulaire ? Grenoble, Cerat. (En ligne sur *Lexicométrica*, 3 (http://www.cavi.univ-paris3.fr/lexicometrica).

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie.* Genève-Paris, Slatkine-Champion.

Lebart L. et Salem A. (1994). *Statistique textuelle*. Paris, Dunod.

Mayaffre D. (2004). *Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la Vème République*. Paris, Champion.

Mayaffre D. (2005). De la lexicométrie à la logométrie. *Astrolabe* (http://www.uottawa.ca/academic/arts/astrolabe).

Mellet S. et Purnelle (2002). Les atouts multiples de la lemmatisation : l'exemple du latin. In *JADT 2002*. Saint-Malo, IRISA-INRIA, Vol 2 : 529-539.

Mellet S. (2003). Lemmatisation et encodage grammatical : un luxe inutile ? *Lexicometrica*, numéro spécial (http://www.cavi.univ-paris3.fr/lexicometrica/thema/thema1/spec1-texte2.htm)/.

Monière D. Labbé C et Labbé D. (2005). Les particularités d'un discours politique : les gouvernements minoritaires de Pierre Trudeau et de Paul Martin au Canada. *Corpus* 4 : 79-101.