

Le sens des mots. L'Europe dans le vocabulaire de Jacques Chirac

Cyril Labbé, Dominique Labbé

▶ **To cite this version:**

Cyril Labbé, Dominique Labbé. Le sens des mots. L'Europe dans le vocabulaire de Jacques Chirac. Documents Numériques, A paraître. halshs-02299918v1

HAL Id: halshs-02299918
https://shs.hal.science/halshs-02299918v1

Submitted on 28 Sep 2019 (v1), last revised 7 Dec 2019 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Le sens des mots

L'Europe dans le vocabulaire de Jacques Chirac

CYRIL LABBÉ
UNIVERSITÉ GRENOBLE ALPES, CNRS, GRENOBLE INP, LIG,
F-38000 GRENOBLE, FRANCE
cyril.labbe@imag.fr

DOMINIQUE LABBÉ
UNIVERSITE GRENOBLE ALPES, PACTE
F-38000 GRENOBLE, FRANCE
dominique.labbe@umrpacte.fr

RESUME.
Présentation du corpus des interventions de J. Chirac durant ses deux présidences (1995-2002 et 2002-2007) : poids des différents modes de communication, description du vocabulaire, vocables les plus utilisés. On constate une stabilité de la communication sur les 12 ans et un poids considérable de l'Europe dans les principaux thèmes. Mais quel sens donnait J. Chirac à ce mot ? Pour répondre à cette question, on reconstitue le réseau d'associations, d'oppositions et de substitutions qui relie ce mot aux autres vocables du corpus. Les principales caractéristiques de cet univers montrent que le président Chirac avait une attitude distanciée vis-à-vis de la "construction de l'Europe".

ABSTRACT.
Presentation of the corpus of the interventions by Jacques Chirac during his two presidencial terms (1995-2002 and 2002-2007): the different modes of communication, description of the vocabulary, especially the most used words. There is a stability of J. Chirac's communication over the 12 years and a considerable weight of Europe in the main themes. But what meaning did J. Chirac give to this word? To answer this question, we reconstruct the network of associations, oppositions and substitutions that links this word to the other terms of the corpus. The main characteristics of this universe show that President Chirac had a distanced attitude towards the "construction of Europe".

MOTS-CLES : France, président, discours politique, Chirac, Sens des mots, Europe, Lexicométrie, Statistique lexicale, Univers lexical

KEYS WORDS : France, president, Chirac, Political discourses, Word meanings, Europe, Lexicometry, Lexical statistics, Lexical universes

Manuscrit des auteurs

Article à paraître dans la revue *Document Numérique* (Lavoisier)

1. Introduction

Avec la numérisation de grandes masses de textes et leur mise en ligne, l'analyse du discours politique peut reprendre de nombreuses questions jusqu'ici sans réponse satisfaisante. Parmi celles-ci : le sens des mots. En effet, la politique est grande consommatrice de mots affectivement chargés et massivement polysémiques (Fabre et al., 1997) : démocratie, liberté, nation, patrie, peuple… Mais quelles significations les hommes politiques leur donnent-ils ?

Depuis Saussure (1916), il existe un relatif consensus autour de l'idée selon laquelle le sens des mots leur vient du lexique de la langue et de l'usage particulier qu'en font les locuteurs de cette langue. Si l'on accepte de voir dans la nomenclature des dictionnaires de langue, une image de la structure du lexique de cette langue – image certes imparfaite mais partagée par tous les usagers de cette langue –, la première tâche de l'ordinateur sera, pour chaque mot, de retrouver son entrée de dictionnaire. Quant à étudier les usages de ce mot, il est proposé d'utiliser des collections de textes rassemblés de manière raisonnée et constituées en corpus.

Nous allons illustrer ces deux procédures à l'aide d'un exemple : les interventions de J. Chirac pendant sa présidence[1]. Ces interventions n'ont jamais été dépouillées exhaustivement et il existe fort peu d'études portant sur le discours de ce président qui a pourtant joué un rôle de premier plan dans la politique française pendant plus de 30 ans.

Né en 1932, J. Chirac a été proche collaborateur du Premier ministre Pompidou (1962-1967) puis député de la Corrèze (1967-1995), dont il préside le conseil général (1970-79). Il a également été maire de la capitale (1977-1995). Il entre au gouvernement en 1968 et il est ministre de l'intérieur au moment du décès de G. Pompidou (avril 1974), ce qui lui donne un rôle clef lors de l'élection présidentielle et lui permet de devenir Premier ministre de 1974 à 1976 puis de prendre le contrôle du parti gaulliste (RPR), contrôle qu'il conserve jusqu'en 1994. Il est une nouvelle fois Premier ministre de 1986 à 88. Enfin, après avoir échoué deux fois à la présidentielle (1981 et 1988), il est élu président en mai 1995 et réélu en mai 2002. Il reste ainsi à la tête de l'Etat pendant 12 ans. Il a été victime d'un grave accident vasculaire cérébral en septembre 2005.

Durant ses deux mandats, il est intervenu près de 2 500 fois. La présentation de ce corpus montrera l'importance du mot Europe. Un calcul permettra de reconstituer l'univers lexical de ce mot et de préciser ainsi le sens que J. Chirac lui associait.

---

[1] Il n'existe pas d'ouvrages de référence sur la présidence Chirac. On peut se reporter à ses mémoires (Chirac, 2009), au livre de Péan (2007) et pour la politique étrangère, dont il sera plus particulièrement question dans cet article, à l'ouvrage collectif dirigé par Lequesne et Vaïsse (2012) et au mémoire de Deligne (2013).

2. Les corpus

Le corpus est une collection systématique de textes ayant subi une série de traitements qui vont être succinctement décrits. Un certain nombre d'outils informatiques permettent d'analyser ces corpus, notamment de rechercher l'univers lexical d'un vocable.

2.1. Les discours présidentiels français

Du 1er janvier 1959 (instauration de la Ve République française) à mai 2017, sept présidents se sont succédés à la tête de l'Etat : Charles de Gaulle (1959-1969), Georges Pompidou (1969-1974), Valéry Giscard d'Estaing (1974-1981), François Mitterrand (1981-1995), Jacques Chirac (1995-2007), Nicolas Sarkozy (2007-2012), François Hollande (2012-2017). Toutes les interventions disponibles de ces présidents ont été collectées et traitées pour constituer autant de corpus qui permettent de tracer le portrait de chacun d'eux mais aussi celui de la communication présidentielle au cours du dernier demi-siècle (Arnold et al., 2016).

Ces textes sont rassemblés dans une bibliothèque électronique comptant actuellement 25 800 textes, soit plus 62 millions de mots. 40% de cette bibliothèque sont constitués de discours politiques français, québécois et canadiens que D. Labbé et D. Monière collectent depuis plus de trente ans. Au sein de cet ensemble, les discours présidentiels (mai 1958 - mai 2017) représentent 9 872 textes et 16 795 660 mots. Avant d'entrer dans la bibliothèque, chaque texte a été traité de la manière suivante.

Une fiche indique l'auteur, sa fonction, les dates et lieux d'émission ainsi que la nature du texte (par exemple, pour le discours politique : allocution, entretien, conférence de presse, message…). La fiche comporte également la source du document, la date du traitement et le nom de l'opérateur.

A l'intérieur du texte, des balises isolent le texte du "para-texte" (par exemple : les questions des journalistes), afin de ne traiter que les propos de l'auteur étudié, tout en conservant le "para-texte" qui doit être fourni à toute personne consultant la bibliothèque.

Puis, chaque mot du texte est doté d'une étiquette comportant sa "graphie standard" (opération importante pour les noms propres) et son "entrée de dictionnaire" (mot "vedette" et catégorie grammaticale). Par exemple, le féminin et le pluriel d'un adjectif sont groupés sous le masculin singulier de celui-ci, ou encore toutes les flexions d'un même verbe sont groupées sous l'infinitif, etc. Ces conventions ont été présentées par C. Muller (1963 et 1977) et ont été implémentées sur ordinateur (Labbé, 1990). Elles épousent, au plus près, les conventions en usage dans la lexicographie française et sont confiées à des automates qui réalisent la quasi-totalité de l'étiquetage. Il est important de souligner que cet étiquetage est sans erreur (par rapport aux conventions retenues) et que toutes ces informations s'ajoutent au texte proprement dit auquel on ne touche pas.

Pour les présidents, une même nomenclature a été appliquée à l'ensemble de leur communication.

Un premier sous-corpus groupe toutes les interventions intégralement diffusées par la radio et (ou) la télévision : des allocutions et des entretiens – parfois très brefs lorsqu'il s'agit du journal télévisé, parfois d'une heure ou plus – et des conférences de presse tenues à l'Elysée et intégralement télévisées que le Général et ses deux successeurs tenaient à peu près deux fois par an mais qui se sont espacées depuis 1986.

Les autres interventions – non intégralement radio-télévisées – sont classées en quatre sous-corpus : les allocutions, les entretiens, les conférences de presse et les messages de toute nature mais ayant pour point commun d'être des documents signés par le président et diffusés par l'Elysée : communiqués, lettres, articles de journaux, préfaces d'ouvrages…

2.2. Le corpus Chirac

Au cours de ses douze ans à la tête de l'Etat, J. Chirac est intervenu 2 478 fois, soit au total 4 081 700 mots et 24 054 vocables différents. Que représentent ces masses ? Par exemple, les trois tomes des *Mémoires de guerre* du général de Gaulle comptent 372 664 mots ; les plus longs romans en français : *Les Misérables* (Hugo), 564 301 mots ; *les Mystères de Paris* (E. Sue), 578 933 mots. L'édition originale de *A la recherche du temps perdu* (Proust), chez Gallimard (1913-1927) compte 11 volumes de 300 pages en moyenne pour 1 327 859 mots. Si l'on imprimait toutes les interventions du président Chirac au même format, il faudrait donc 34 volumes.

Pourtant, il a été un peu moins bavard que V. Giscard d'Estaing ou F. Mitterrand et surtout, en moyenne annuelle, il a moins parlé que N. Sarkozy et F. Hollande. La comparaison est impossible avec C. de Gaulle et G. Pompidou (dont l'activité a été limitée dans les deux dernières années du fait de sa maladie), puisqu'on ne dispose d'aucune liste exhaustive de leurs interventions.

Chacun des 2 478 textes est doté d'un index décrivant son vocabulaire. Les vocables y sont classés par ordre alphabétique avec toutes leurs flexions attestées et le nombre de leurs occurrences. Un index est également constitué pour chaque sous-corpus et pour le corpus total. Ce dernier comporte 57 357 lignes (les 24 054 vocables avec leurs flexions) et 2 478 colonnes (les textes). Le tableau 1 donne quelques extraits de cet index récapitulatif[2].

Les exemples du tableau 1 ne sont pas anecdotiques. Dans tout texte en français, *être* et *avoir* sont les verbes les plus fréquents et plus d'un mot sur trois peut être rattaché à plus d'une entrée de dictionnaire (homographies). D'ailleurs, imagine-t-on qu'on réponde à un politologue qu'il

[2] L'index alphabétique complet est consultable en ligne sur le site du Centre de Linguistique de Corpus (Université de Neuchatel).

ne peut étudier le "pouvoir" parce que les occurrences du substantif sont mélangées à celles de l'infinitif du verbe homographe ?

Ces index fournissent une porte d'entrée indispensable pour explorer les corpus et retrouver toutes les attestations d'un vocable sous ses diverses flexions. Ils sont également la matière première pour un grand nombre d'analyses.

Tableau 1. Extraits de l'index alphabétique des interventions du président J. Chirac (colonne des totaux)

avion (n. m.)	*230*
avion	*157*
avions	*73*
avoir (v.)	*75 946*
avions	*553*
est (n.m.)	*194*
Est	*147*
est	*47*
être (v.)	*113611*
est	*56 072*
été	*8 131*
être	*10 031*
soit	*3 262*
sommes	*3 093*
suis	*3 937*

être (n. m.)	*135*
être	*5*
pouvoir (v)	*144 083*
peux	*629*
pouvoir	*993*
puis	*71*
pouvoir (nm)	*976*
pouvoir	*406*
pouvoirs	*570*
soit (cj)	*381*
somme (nf)	*103*
somme	*45*
sommes	*68*
suivre	*577*
suis	*12*

2.3. Principales caractéristiques du corpus Chirac

En moyenne, chaque année, le président Chirac est intervenu 207 fois. Si l'on se souvient qu'il prenait un bon mois de vacances durant l'été, une semaine à Noël et à Pâques et ne parlait pas le jour de l'an, à Noël ou à Pâques, etc. cette moyenne signifie que, durant ses jours d'activité, il y avait au moins une intervention. Chaque année, il a prononcé en moyenne 340 142 mots (soit presque l'équivalent des *Mémoires de guerre* du général de Gaulle parues en trois tomes).

Seul le président américain J. Carter, aurait dépassé cette intensité, avec près d'une intervention par jour mais, en moyenne, ces interventions étaient plus brèves (nous devons cette indication à J. Savoy).

Il faut ensuite considérer le genre de l'intervention (tableaux 2 et 3). La radio-télévision ne représente qu'une intervention sur 20 et 8% des mots. Naturellement, ces interventions avaient un auditoire plus vaste mais il serait erroné de réduire la communication présidentielle à ces prestations.

Dans la majorité des cas, le président a sorti un papier de sa poche et a prononcé un discours (ce que C. de Gaulle appelait "inaugurer les chrysanthèmes"), la moitié du temps à Paris et en Ile de France, le reste lors de ses multiples déplacements en province, dans les départements d'outre-mer et à l'étranger. La moyenne (près de 1 900 mots) correspond à environ vingt minutes de parole et l'exercice ne dépasse habituellement pas la demi-heure. Au total pour les douze années et pour ce seul exercice, cela représente plus de 220 heures de parole.

Tableau 2. Les interventions de J. Chirac classées par genre (1995-2007)*

Sous-corpus	Textes	Mots (N)	Vocabulaire (V)
Radiotélévision	117	341 383	7 384
Autres :			
Allocutions	1 169	2 203 256	20 745
Conférences de presse	534	963 992	10 407
Entretiens	131	288 804	7 611
Messages	527	284 265	8 837
Corpus	2 477	4 081 700	24 054

-* y compris la campagne présidentielle à la fin du premier mandat (février – mai 2002)

Tableau 3. Les interventions du président Chirac classées par genre (structure et moyennes)

Sous-corpus	Textes %	Mots (N) %	Longueur moyenne (mots)
Radiotélévision	4,7	8,4	2 918
Autres :			
Allocutions	47,2	54,0	1 885
Conférences de presse	21,6	23,6	1 805
Entretiens	5,3	7,1	2 205
Messages	21,3	7,0	539
Corpus	100,0	100,0	1 647

La plus grande partie des conférences de presse ont été tenues lors des déplacements à l'étranger, soit visites d'état, soit en marge du conseil européen de Bruxelles et des sommets internationaux (G7-G20). Leur relative brièveté est également remarquable.

A part l'allocution, l'entretien était le mode de communication préféré de J. Chirac (d'ailleurs la majorité des interventions à la radio-télévision sont des entretiens).

Au sein de ce vaste ensemble, existe-t-il des solutions de continuité et des ruptures ?

3. Stabilité de la communication et du vocabulaire usuel ?

Les deux mandats de J. Chirac ont connu de nombreux avatars, à commencer par un mouvement social de grande ampleur en nov.-déc. 1995, une dissolution ratée suivie d'une cohabitation de cinq années (1997 – 2002) avec un Premier ministre socialiste (L. Jospin) et un gouvernement de gauche, une campagne présidentielle (2002), un échec au référendum sur la constitution européenne (mai 2005) et de nombreuses crises extérieures (attentats du 11 septembre 2001, guerres dans l'ex-Yougoslavie, en Afghanistan puis en Irak, occupation israélienne du Liban, assassinat du premier ministre libanais R. Hariri, ami personnel du président), etc. On peut donc s'attendre à ce que la communication présidentielle ait évolué dans sa forme, son volume et surtout ses thématiques. Pourtant, cette communication a été relativement stable du moins jusqu'en 2005.

3.1. Stabilité de la communication ?

Pour donner une mesure de cette stabilité d'ensemble, les deux mandats sont comparés en tenant compte de ce que le second n'a duré que cinq ans contre sept pour le premier. Les chiffres absolus sont donc convertis en proportions (tableaux 4 et 5).

Tableau 4. Comparaison des volumes moyens de la communication de J. Chirac entre son premier mandat (1995-2002) et son second (2002-2007)

Moyennes	1 (1995-2002)	2 (2002-2007)	Variation %
Interventions par an	186	234	+ 26 %
Mots par an	349 265	327 368	- 6 %
Longueur (en mots)	1 874	1 395	- 26 %

Entre le premier et le second mandat, on constate une augmentation d'un quart du nombre des interventions mais une réduction équivalente de leurs longueurs qui se traduit d'ailleurs par un léger recul du volume moyen de mots émis chaque année. Ces évolutions sont à mettre en relation avec celles des différents genres (tableau 5).

Tableau 5. Comparaison de la communication de J. Chirac classée par genres entre ses deux mandats (1 : 1995-2002 ; 2 : 2002-2007)

	Interventions		Mots (N)		Longueur moyenne	
Mandats	1	2	1	2	1	2
Radiotélévision	5,4	3,9	8,1	8,8	2 777	3 135
Autres :						
Allocutions	54,6	38,9	59,9	45,1	2 054	1 620
Conférences de	23,5	19,4	22,8	24,9	1 813	1 795
Entretiens	4,9	5,7	5,2	9,9	1 983	2 416
Messages	11,5	32,1	4,1	11,3	664	490
Total général	100,0	100,0	100,0	100,0	1 873	1 395

Au cours du second mandat, les interventions ont été plus nombreuses mais le président a prononcé un peu moins d'allocutions et celles-ci ont été nettement plus brèves. En contrepartie, il a privilégié les entretiens surtout avec la presse écrite mais aussi avec les radios et télévisions, ce qui explique l'allongement de la longueur moyenne des interventions radio-télévisées ; puis les conférences de presse, en général assez brèves, sont également des formes d'entretiens ; enfin, les messages dont le poids a été multiplié par 2,4 mais qui sont plus brefs qu'auparavant.

Ce glissement a été progressif et s'explique peut-être par la fatigue puis la maladie. Cela est particulièrement évident pour les messages dont la multiplication s'est produite après son accident cérébral (septembre 2005) et durant sa convalescence, comme si l'entourage du président avait tenté de pallier par là son absence ou du moins de la masquer.

En définitive, seules la fatigue et la maladie ont pu altérer une communication bien rodée. Cette relative stabilité semble confirmée par l'examen des vocables préférés.

3.2. Les vocables préférés du président Chirac

Parmi les documents électroniques constituant le corpus, l'index hiérarchique donne les vocables les plus employés. Le tableau 6 présente la tête de liste. Les effectifs absolus sont convertis en fréquence relative exprimée en "pour mille mots".

Tableau 6. Les vocables les plus employés dans les interventions de J. Chirac

Rang	Lemme et Catégorie grammaticale	Effectifs	Fréquence (‰)
1	le (art.	493 884	121,0
2	de (pré.)	347 460	85,1
3	à (pré.)	113 943	27,9
4	être (v.)	113 673	27,8
5	et (conj.)	112 669	27,6
6	un (art.)	78 757	19,3
7	avoir (v.)	75 977	18,6
8	je (pro.)	56 548	13,9
9	qui (pro.)	49 594	12,2
10	que (conj.)	46 338	11,4

Les mots les plus fréquents jouent un rôle essentiellement syntaxique et n'acquièrent un contenu qu'avec le contexte de l'énonciation. Sauf le "je" (qui est toujours le président), le premier mot non-outil – le nom propre France – se trouve au 31e rang avec 16 410 occurrences (soit 4,02 ‰ de la surface des textes). Avec une fréquence inférieure à un demi pour cent, "France" est donc trente fois moins fréquent que l'article "le", bien que beaucoup plus important pour l'analyse de la communication présidentielle.

Dans ce vaste corpus, 41 des 43 vocables d'effectifs supérieurs à 10 000 occurrences sont des mots outils. A eux seuls, ils couvrent plus de la moitié de la surface des textes (53%) bien qu'ils ne constituent que 0,16% du vocabulaire. A l'opposé, 90% des vocables apparaissent moins de 100 fois et ne couvrent que 5,2% de la surface des textes. Ce sont pourtant eux qui véhiculent l'essentiel du message.

Autrement dit, le vocabulaire d'un corpus est une grande collection d'événements rares très inégalement distribués, dont les plus fréquents ne sont pas les plus facilement interprétables. Ce phénomène d'inégale distribution des fréquences a été mis en lumière par Zifp 1935 (voir également Mandelbrot, 1957).

De plus, ces listes ne résolvent pas la question de savoir si l'usage d'un vocable est conjoncturel – c'est-à-dire lié à un événement particulier – ou relativement permanent et que l'on peut donc considérer comme faisant partie du lexique du locuteur. C. de Gaulle en donne un exemple éclairant : entre 1958 et 1962, "Algérie" est le troisième nom propre le plus employé par le Général (après France et Français) mais, après juillet 1962, ce vocable disparaît totalement de son vocabulaire : il n'appartiendrait donc pas au lexique permanent (ou fondamental) du Général (sur la notion de répartition des vocables sur la surface d'un corpus : Labbé et Labbé, 2017).

Ces réserves admises qu'apprennent ces listes organisées par catégories grammaticales ?

Les tableaux 7 et 8 ci-dessous comparent, dans les deux mandats du président Chirac et, pour chacun des mots les plus fréquents, leur rang et le nombre de fois qu'ils ont été employés (effectifs). Comme les deux corpus n'ont pas la même longueur, les effectifs absolus sont convertis en fréquences exprimées en pour mille mots (‰).

Tableau 7. Les dix noms propres les plus fréquents dans les deux corpus (fréquences en pour mille mots)

	1995-2002		2002-2007	
Rang	Vocable	F (‰)	Vocable	F (‰)
1	France	3,88	France	4,23
2	Europe	1,97	Europe	1,88
3	Français	1,08	Français	0,97
4	Union Européenne	0,71	Union Européenne	0,71
5	Afrique	0,46	Afrique	0,66
6	Paris	0,36	Chine	0,39
7	Russie	0,31	Nations Unies	0,38
8	Etats-Unis	0,28	Allemagne	0,37
9	Amérique	0,23	Liban	0,35
10	Asie	0,23	Irak	0,35

Entre les deux mandats, les cinq premiers noms propres se retrouvent dans le même ordre et avec des densités proches. Au-delà du 5e rang, le poids de la conjoncture commence à se faire sentir. A son arrivée au pouvoir, la Russie occupait encore un rang important dans les relations internationales. Puis après septembre 2001, ce fut, pendant quelques mois, les Etats-Unis d'Amérique. Durant le second mandat, le président a fait appel aux Nations-Unies pour prévenir ou résoudre les guerres de Yougoslavie, du Liban et d'Irak.

Tableau 8. Les dix substantifs les plus fréquents dans les deux corpus Chirac

1995-2002			2002-2007	
Rang	Vocable	F (‰)	Vocable	F (‰)
1	pays	3,23	pays	3,13
2	monsieur	2,42	monsieur	2,52
3	monde	1,85	monde	1,93
4	président	1,76	président	1,88
5	état	1,52	développement	1,55
6	problème	1,28	état	1,40
7	homme	1,16	problème	1,34
8	développement	1,08	action	1,08
9	paix	1,03	année	1,03
10	an	0,97	ministre	1,03

Comme pour les mots à majuscules initiales, les premiers substantifs se retrouvent pratiquement dans le même ordre avec des densités voisines. En allongeant la liste jusqu'au centième rang, les deux colonnes ont plus de 80 % des substantifs en commun avec cependant des variations de densités qui peuvent s'expliquer par l'actualité plus que par des changements politiques (tableau 8). Un constat semblable peut être fait avec les verbes, les adjectifs, les pronoms et les adverbes. Cela amène deux conclusions : d'une part, il y avait une stabilité remarquable dans le discours malgré les avatars signalés ci-dessus et, d'autre part, tout au long de ses deux mandats, l'Europe a pesé d'un poids considérable. En ajoutant à ce mot "Union Européenne", c'est le second thème après France. Cette conclusion est renforcée en considérant la famille de mots formant un paradigme avec "Europe" (c'est-à-dire que ces mots peuvent se substituer à "Europe" dans certains contextes).

- "Européen" (nom de peuple) au 18e rang des noms propres : 631 occurrences, soit 0,16 pour mille mots.

- L'adjectif "européen" : 5 633 occurrences (1,38 ‰) au troisième rang des adjectifs juste derrière "grand" et "nouveau" et largement devant l'adjectif "français" (4 982 occurrences, 1,22‰). Près de 500 fois, il s'agit de la "construction européenne" qui est l'une des combinaisons figées les plus utilisées du président.

Il faut encore ajouter 121 occurrences de "union" – sans adjectif derrière, mais désignant explicitement l'"Union Européenne" – et 95 fois "communauté" (également sans adjectif), employée pour "Communauté Européenne" (elle-même utilisée 17 fois), et 16 "Marché Commun". Il faudrait aussi prendre en considération les expressions "marché unique", "grand marché", etc.

Malgré ce flottement dans les dénominations – qui est lui-même révélateur –, il ne fait pas de doute que, durant les 12 ans de présidence Chirac, l'Europe a été un des thèmes principaux

de son discours, après la France mais avant les Français ou l'Afrique à qui le président accordait pourtant beaucoup d'importance, manifestée par le nombre de voyages officiels, et de rencontres avec des chefs d'Etats, presqu'aussi nombreux qu'avec les pays européens. Cette conclusion n'est pas surprenante : dans les discours des présidents français, depuis juillet 1962, l'Europe occupe toujours la deuxième place après la France.

Mais quel sens J. Chirac donnait-il à ces deux mots "Europe" et "Union européenne" ? Cette question découle logiquement du flottement sémantique signalé plus haut. De plus, J. Chirac semble avoir beaucoup varié sur ce point au cours de sa carrière : il a pris des positions hostiles à l'intégration européenne, notamment lors des premières élections du parlement européen au suffrage universel en 1979, avant de se rallier au traité de Maastricht instituant l'Union Européenne (1992) puis de promouvoir, dans le traité de Nice (2000), une Europe-fédération qui culmine avec le projet de constitution (2005) rejetée par les Français (Laurent et Sauger, 2005).

4. Le calcul du sens

Pour retrouver le sens d'un vocable dans un ensemble de textes, l'algorithme doit chercher les mots qui sont associés à "Europe". Cette association peut avoir un triple visage : attirance mutuelle, répulsion ou neutralité. Dans le premier cas, les deux mots ont tendance à apparaître ensemble. A l'inverse, ils peuvent s'exclure mutuellement de manière plus ou moins rigoureuse. Enfin l'apparition de plusieurs mots dans les mêmes contextes signale qu'ils peuvent se substituer entre eux, et qu'ils appartiennent à une même famille de mots (paradigme). Par exemple, dans les propos de J. Chirac, l'"Union européenne" était-elle un substitut possible d'"Europe" et dans quels contextes ?

Le concordancier est l'outil usuel pour retrouver les usages d'un mot dans un corpus (ce programme liste, généralement en une ligne, le contexte avant et arrière de chaque occurrence du mot recherché). Ici l'utilisation de la concordance serait difficile car elle comporte, pour "Europe", 7 896 lignes et pour "Union Européenne" : 2 914 lignes. Il faut donc avoir recours à des outils capables de donner des réponses plus synthétiques.

4.1. Le calcul de l'univers lexical d'un vocable

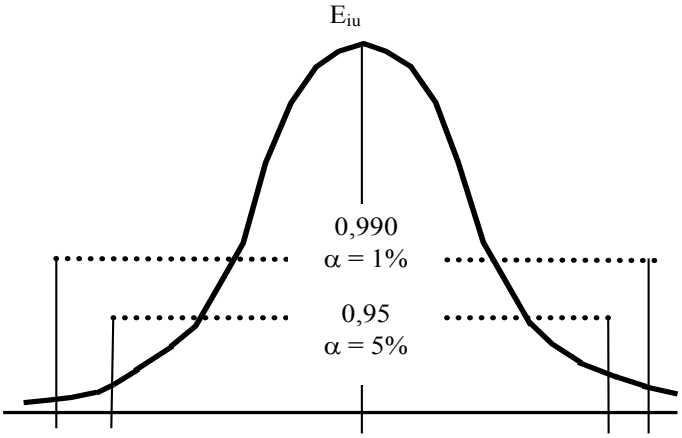
Dans un corpus, le sens particulier d'un vocable est donné par l'ensemble des relations d'attirance, de répulsion et de substitution qui le lient aux autres vocables de ce corpus. (Le calcul de ces relations a été présenté dans Hubert et Labbé, 1995 ; Leselbaum et Labbé, 2002 ; Labbé et Labbé, 2005 ; Labbé, 2010 ; pour une application sur un corpus anglais : Arnold, 2008).

Le calcul est inspiré des "spécificités" utilisées en analyse des données textuelles (Lafon, 1980 et 1984) mais, outre qu'il porte sur les vocables (et non les formes graphiques), il s'en

sépare sur plusieurs points exposés ci-dessous. C'est pour rappeler ces différences que l'on parle de "vocabulaire caractéristique" (au lieu de "spécifique").

L'algorithme extrait du corpus entier toutes les phrases qui contiennent le vocable recherché. Ce sous-corpus est l'univers du vocable. Le programme établit le vocabulaire de cet univers et le compare à celui du corpus entier, en testant pour chaque vocable, l'hypothèse (H_0) selon laquelle sa densité dans l'univers ne diffère pas significativement de celle dans le corpus entier, en acceptant une marge d'incertitude α (souvent qualifiée de "marge d'erreur") de 5% ou 1% (Graphique 1).

Graphique 1. Schéma de principe des intervalles d'acceptation ou de rejet de H_0



Ce schéma de principe rappelle d'abord que, dans une vaste population homogène, tout caractère statistique est soumis à des fluctuations aléatoires autour de sa moyenne, les valeurs se répartissant selon une courbe en cloche qui délimite des zones d'acceptation ou de rejet de H_0 selon α (traits pointillés).

Avec

N : nombre de mots dans le corpus (ou longueur),

V : le nombre de vocables différents dans le corpus (ou vocabulaire),

U : l'univers d'un vocable inclus dans le corpus,

N_u et V_u : respectivement le nombre de mots et de vocables dans l'univers U,

et $N >> N_u$

Soit un vocable i d'effectif F_i dans le corpus total et F_{iu} dans l'univers. Si la même loi de distribution gouverne les occurrences de i dans le corpus entier et dans U, le nombre de ses apparitions dans U doit se situer dans la plage de fluctuation normale autour de :

$$E_{iu} = F_i * \frac{N_u}{N} \quad (1)$$

Si la valeur observée (F_{iu}) s'écarte de cette valeur attendue, se situe-t-elle à l'intérieur ou à l'extérieur de la plage de fluctuation normale ? Autrement dit, considérant les paramètres N, N_u et F_i, quelle est la probabilité pour que surviennent l'événement F_{iu} ? Cette probabilité (P) est calculée grâce à la formule (2) (loi hypergéométrique : Labbé et Labbé 1995) :

$$P(X = F_{iu}) = \frac{\begin{bmatrix} F_i \\ F_{iu} \end{bmatrix}\begin{bmatrix} N-F \\ N_u-F \end{bmatrix}}{\begin{bmatrix} N \\ N_u \end{bmatrix}} \quad (2)$$

On en tire un indice L qui mesure la surface de la courbe qu'il faut parcourir pour atteindre la valeur F_{iu} (en la parcourant de gauche à droite, en partant de zéro) :

$$L = \sum_0^{F_{iu}} (P(X \leq F_{iu})) \quad (3)$$

Les bornes sont les suivantes :

- L < 0,005 ou L < 0,025 : on a parcouru moins de 0,5 % (ou moins de 2,5 %) de la surface de la courbe : l'observation se trouve à l'extérieur de l'intervalle et à gauche de celui-ci. Il y a moins de une chance (ou 5 chances) sur 100 de se tromper en affirmant que la valeur observée est significativement inférieure à la valeur attendue (H_0 est rejetée). On considère que le vocable est une caractéristique négative (C-) de l'univers.

- Pour un indice supérieur à 0,025 et inférieur à 0,975 : H_0 ne peut être rejetée et l'on considère que la fréquence d'emploi dans le corpus et dans l'univers ne diffèrent pas : le vocable est supposé commun au corpus et à l'univers et sera noté C_0 pour "non-caractéristique" ou "conforme à l'hypothèse nulle".

- Pour un indice supérieur à 0,975 ou 0,995 il y a respectivement moins de 5 % ou 1 % de chances de se tromper en affirmant que la valeur observée est significativement supérieure à la valeur attendue (H_0) et que le vocable i est une caractéristique positive (C+) de U.

4.2. Remarques

L'indice est calculé sur les vocables et non sur les formes graphiques (comme le font les calculs de spécificités dans les logiciels d'analyse des données textuelles). Par exemple, le calcul sur les formes graphiques donne comme plus fortes spécificités associées à "Europe" : "l'", "d'", "une", "est" ou encore "forte", et comme plus fortes spécificités négatives :"la", "les", "de", "un", "suis" et "sont", ou "fort" et "forts". Autrement dit, le calcul confirme que J. Chirac élide l'article devant une voyelle et accorde l'article ou l'adjectif avec le substantif ou encore le verbe avec le sujet. En revanche, une analyse sur les formes graphiques ne permet pas de savoir si, lorsqu'il parlait de l'Europe, J. Chirac préférait l'article "le" ou "un", le verbe "être", l'adjectif "fort", etc. Autrement dit, pour atteindre le niveau sémantique, il faut d'abord

neutraliser le "bruit" engendré par les composants graphique et syntaxique de la langue, ce que permettent la standardisation des graphies et l'étiquetage des mots.

L'indice fournit directement l'information souhaitée. Il mesure la force de la liaison (attirance ou répulsion) existant entre un vocable i et celui dont on recherche l'univers.

Le test est bilatéral (pour chaque vocable les situations C+ et C- sont possibles), c'est pourquoi les valeurs seuils sont déterminées avec $\alpha/2$. En cas de distribution normale, seulement 5 % ou 1 % des vocables devraient être caractéristiques et réparties également entre C+ et C-.

Le vocabulaire C_0 n'a pas le même statut que C+ et C-. Le fait qu'on ne puisse rejeter H_0 ne signifie pas qu'elle est vraie car l'on ne peut y associer un risque (dit de "deuxième espèce"). Pour limiter celui-ci, on travaillera avec l'intervalle suivant : $0{,}05 < L < 0{,}95$. En cas de distribution normale, cet intervalle devrait contenir 90 % du vocabulaire.

4.3. Trois compléments

Ce calcul se heurte à quelques difficultés généralement négligées en analyse des données textuelles.

4.3.1. Lien avec la catégorie grammaticale

Si un vocable appartient à une catégorie sur-employée dans U, il a de bonnes chances d'être C+ et vice-versa (voir Monière et Labbé, 2012). Par exemple, le verbe "être", qui est le plus fréquent dans tout texte en français. Il a pour fonction essentielle d'exprimer une situation, un état, une qualité. Il est donc important de savoir si J. Chirac utilisait plus ou moins cette fonction quand il parlait de l'Europe.

Ce verbe apparaît 113 673 fois dans le corpus entier et 6 162 fois dans l'univers. Dans le calcul classique, l'effectif attendu dans U est :

$$E_{iu} = F_i * \frac{N_u}{N} = 113\,673 * \frac{230\,105}{4\,081\,700} = 6\,409$$

Or le verbe "être" n'apparait que 6 162 fois dans U, soit 247 occurrences de moins que ce que laisserait attendre H_0. La formule (3) indique que cet écart a moins de 3 chances sur 10 000 de se produire par hasard.

Dans le vocabulaire de J. Chirac, le verbe "être" serait donc une caractéristique négative de l'Europe ?

Cependant, lorsqu'il parlait de l'Europe, J. Chirac employait 9% de verbes en moins par rapport à sa propension moyenne lorsqu'il traitait d'autres sujets (Annexe 1). Dès lors, la question devient : parmi les verbes utilisés pour traiter de l'Europe, "être" était-il privilégié ou

sous-utilisé ? Dans le corpus entier, les verbes représentent 569 101 mots (NV) et 29 350 (NV_u) dans l'univers. Dans les formules (1) et (2) on remplace N par NV et N_u par NV_u. Considérant l'effectif de "être" dans NV, H_0 en laisse attendre dans U :

$$E_{iu} = F_i * \frac{NV_u}{NV} = 113\,673 * \frac{29\,350}{569\,101} = 5\,862$$

Or le verbe "être" apparait 6 162 fois dans U, soit 300 occurrences de plus qu'attendu. La formule (3) indique que cet écart a moins de 1 chance sur 10 000 de se produire par hasard. Le verbe "être" est donc une caractéristique positive de l'Europe. Ou encore, lorsqu'il parlait de l'Europe, J. Chirac avait une propension moyenne à utiliser le verbe "être" significativement plus forte que lorsqu'il traitait d'autres sujets.

Cet exemple n'est pas anecdotique : parmi les verbes que cette rectification fait passer de C- à C+, on trouve : "faire" – dans un texte en français, ce verbe occupe toujours le troisième rang derrière "être" et "avoir" et c'est le principal verbe d'action – "vouloir", "pouvoir", "aller", "souhaiter", "décider". Il n'est pas indifférent de savoir que J. Chirac associait à l'Europe, les modalités de la volonté, du possible, du souhait ou du futur indéterminé (qu'exprime le pseudo-auxiliaire "aller") (sur ces modalités, voir Labbé et Labbé, 2010).

Toutes les catégories grammaticales sont concernées par ce problème, notamment les substantifs, adjectifs et pronoms. Par exemple, le pronom "on" passe lui aussi de C- à C+. Chez J. Chirac, ce pronom désignait "des personnes que je ne nomme pas et qui critiquent ou s'opposent à ma politique". Il surgissait à chaque fois que J. Chirac répondait à l'opposition sur des points sensibles… comme l'Europe.

A l'inverse, beaucoup de substantifs ou d'adjectifs passent de C+ à C-. C'est surtout vrai pour les noms propres (leur densité est multipliée par 2,5 lorsque J. Chirac parlait de l'Europe). Le cas le plus frappant est "Afrique" (5e nom propre le plus utilisé par J. Chirac au cours de ses deux mandats). Certes, le président ne manquait pas de demander l'aide économique et financière de l'Europe, mais l'Afrique n'en restait pas moins un vocable sous-employé quand il parlait de l'Europe sans doute parce que, dans son esprit, elle demeurait la chasse gardée de la France.

En conclusion sur ce premier point, pour donner à chaque vocable la même chance d'être C+, C- ou C_0, il faut pondérer ses effectifs par le poids de sa catégorie grammaticale dans le corpus et dans l'univers. Et, pour faciliter l'interprétation des résultats du calcul il faut également ventiler ces résultats par catégories grammaticales.

4.3.2. Lien avec la fréquence d'emploi.

Le calcul des probabilités comporte un postulat : plus il y a d'observations plus elles convergent vers la valeur moyenne (ou "modale" : la plus nombreuse). En statistique lexicale, cette convergence existe bien mais elle est moins rapide que ce que laisserait attendre le modèle

probabiliste. De ce fait, plus un vocable est fréquent, plus il a de chances d'être C+ ou C- (voir Monière et Labbé, 2018).

Pour neutraliser ce phénomène, le logiciel range les vocables caractéristiques en trois classes : peu fréquents, fréquents, très fréquents. Les bornes de ces classes sont déterminées automatiquement par le logiciel pour ventiler également, dans ces classes, les vocables compris dans les calculs. Par exemple, dans V, il y a V' verbes dont les effectifs sont suffisants pour qu'ils soient susceptibles d'être C+ ou C-. Les bornes sont déterminées pour contenir les 0,33V' moins fréquents, puis les 0,33V' suivants (fréquents), enfin le dernier tiers des plus fréquents. Le logiciel donne à chacune de ces classes le même poids dans les listes du vocabulaire caractéristique. Cette opération neutralise en bonne partie l'influence de la fréquence et surtout, elle permet de retrouver les vocables rares mais caractéristiques qui, dans le calcul classique, sont noyés parmi les vocables très fréquents. Par exemple, sans ce procédé, il serait impossible de retrouver "la forteresse Europe" (expression employée 12 fois alors que H_0 en laisse attendre moins de 1) ou "le leadership de la France" (employée 7 fois contre 0,5 attendus).

4.3.3. Questions de seuils

Tous les vocables de V ne peuvent entrer dans le calcul. Celui-ci ne peut porter que sur les vocables dont les effectifs sont supérieurs à certains seuils.

Seuil de liaison positive : quel est l'effectif minimum dans le corpus total pour qu'un vocable soit susceptible d'être C+ lorsque toutes ses occurrences surviennent dans U ($F_i = F_{iu}$) ?

Pour que i soit C+ (avec $\alpha = 1$), il faut que : $F_i \geq 5$ (avec $F_i = F_{iu}$ et $N >> N_u$ => $E_{iu} = \varepsilon$)

Seuil de liaison négative : quel est l'effectif minimum pour qu'un vocable soit susceptible d'être C- lorsqu'aucune occurrence ne survient dans U ? C'est-à-dire quelles sont les conditions pour qu'on obtienne un indice inférieur à 0,025 ? Il faut pour cela que l'effectif attendu (avec H_0) soit au moins de 5 :

Avec $F_{iu} = 0$; $N >> N_u$ et $\alpha = 1$; $E_{iu} \geq 5$

Rappel :$E_{iu} = F_i * N_u/N$

=> Seuil C- : $5*N/N_u$

Soit par exemple avec N_u = 5% de N, ce seuil sera $F_i = 5 * 100/5 = 100$

Détermination du vocabulaire commun (C_0). En toute logique le seuil C- s'applique aussi à la détermination du vocabulaire commun à l'univers et au corpus. En effet, pour pouvoir accepter H_0, il faut que la fréquence du vocable soit telle qu'il puisse être C+ et C- et que l'indice se situe entre les deux bornes définies ci-dessus. Dans ce cas, seule une petite proportion du vocabulaire du corpus sera susceptible d'appartenir au vocabulaire commun. Si l'on souhaite élargir cette liste, le seuil choisi devra tenir compte de deux conditions minimales. D'une part,

le vocable considéré doit avoir une densité telle dans le corpus entier (F_i) que sa fréquence théorique dans U soit au moins égale à 1 ($E_{iu} \geq 1$) et, d'autre part, que sa présence dans l'univers soit effectivement avérée ($F_{iu} \geq 1$). Il serait d'ailleurs raisonnable de considérer qu'on ne peut tirer aucune conclusion générale d'un cas unique. La condition minimale devient alors $F_{iu} \geq 2$.

En tout état de cause, les calculs ne peuvent porter que sur les vocables dont le nombre d'occurrences est au moins de 5 dans le corpus total. De plus, si l'on veut donner aux vocables compris dans ce calcul, une chance égale d'être C+ ou C- ou C_0, le seuil devra être beaucoup plus élevé et le calcul ne concernera qu'une très faible proportion du vocabulaire, proportion composée des mots les plus fréquents qui ne sont pas forcément les plus intéressants.

On retrouve ainsi la principale difficulté de toute statistique appliquée au langage : le vocabulaire est composé d'un grand nombre d'évènements très rares. Le calcul des univers n'est donc pertinent qu'avec de grands corpus et sur des vocables d'effectifs importants. Tel est le cas du corpus Chirac et du vocable "Europe".

5. Le vocabulaire caractéristique de l'Europe chez J. Chirac

Appliqué au corpus Chirac, ce calcul permet de mettre au jour les vocables associés à "Europe" (positivement et négativement) mais aussi ceux employés autant à propos de l'Europe que pour les autres thèmes.

5.1. Principales caractéristiques de l'univers de "Europe"

Rappelons les caractéristiques de ce corpus : 4 081 700 mots (N) et 24 054 vocables (V)

L'univers de "Europe" contient : 230 105 mots (N_u) et 6 285 vocables (V_u)

$N_u / N = 5{,}64\%$ donc : $N >> N_u$

Dans ce cas, avec $\alpha = 1$, le seuil minimal pour qu'un vocable soit C+ est $F_{iu} > 4$; le seuil minimal pour qu'un vocable soit C_0 est $F_{iu} > 1/0{,}0564 > 18$; le seuil minimal pour qu'un vocable soit C- est $F_{iu} > 4 * (1/0.0564) > 70$.

En ne considérant que le seuil C+, le nombre de vocables susceptibles d'être caractéristiques est 10 506 (sur un total de 24 053), soit 43,7% du vocabulaire total. Il y a 1 145 vocables effectivement caractéristiques de l'univers, soit 10,9 % des vocables du corpus ayant une fréquence suffisante pour entrer dans le calcul mais ces vocables représentent 81 % de la surface totale de l'univers. Cette différence confirme que les vocables les plus fréquents sont aussi les plus nombreux à être caractéristiques.

Cependant, cette surface totale peut être trompeuse car il ne faut pas raisonner en termes de présence/absence – la plupart de ces vocables étant présents à la fois dans l'univers et dans le

reste du corpus – mais en termes d'écarts entre les observations et ce que donnerait l'hypothèse H_0 (E_{iu}). Pour un vocable i, cet écart absolu sera :

$$Ecart = |F_{iu} - E_{iu}|$$

Il y a 548 vocables C+ (soit 8,7% du vocabulaire de l'univers). La somme de leurs écarts par rapport à H_0 est de 23 054 mots, soit 10 % de la surface de l'univers.

Il y a 597 vocables C- (soit 9,5% du vocabulaire de l'univers). La somme de leurs écarts à H_0 représente 16 708 mots que l'on peut considérer comme manquants dans l'univers (soit 7,3 pour cent de sa surface).

L'asymétrie dans les sommes des écarts pour les C+ et les C- s'explique par les questions de seuil évoquées dans la présentation du calcul.

Au total, la somme de ces écarts représente 39 764 mots soit 17% de l'univers alors que l'hypothèse H_0 en laissait attendre 1%. On en conclut donc que le vocabulaire utilisé quand le président parlait de l'Europe est nettement décalé par rapport à celui employé pour les autres thèmes.

Pour chacun des vocables dont les effectifs sont supérieurs aux seuils choisis, l'algorithme calcule son éventuelle liaison avec "Europe" puis classe ces vocables par indices décroissants. Les tableaux 9 à 11 donnent des extraits de ces résultats : premiers vocables C+, puis C- et enfin les C_0 (l'annexe 2 donne des listes plus complètes).

5.2. Les principales liaisons positives avec "Europe"

Quels sont les vocables les plus fortement attirés dans l'orbite d'"Europe" ? Le tableau 9 donne les dix premiers avec le détail du calcul.

Tableau 9. Les vocables les plus caractéristiques de l'Univers de "Europe" dans les propos du président Chirac (classement en fonction de la force de la liaison)

Rang	Vocable	Catégorie gram.	Effectif corpus	Effectif univers	Effectif théorique	Proportion	Indice
1	renforcer	v	1 918	203	98,9	0,106	1
2	modèle	n m	835	110	40,5	0,132	1
3	main	n f	699	94	33,9	0,134	1
4	élargir	v	340	71	17,5	0,209	1
5	alliance	n f	641	93	31,1	0,145	1
6	euro	n m	981	113	47,6	0,115	1
7	pacifique	adj	303	69	16,7	0,228	1
8	monde	n m	7 676	852	372,4	0,111	1
9	Etats-Unis	n p	1 224	255	75,7	0,208	1
10	demain	adv	1 662	373	90	0,224	1

Le tableau se lit de la manière suivante : le verbe "renforcer" apparaît 1 918 fois dans le corpus entier et 203 fois dans les phrases contenant "Europe", alors que l'on en attend 99. Ces 203 occurrences représentent 10,6% du total des emplois de ce verbe alors qu'un emploi normal en laisserait attendre une proportion équivalente à celle que pèse l'univers (5,64%). Il y a moins d'une chance sur 10 000 de se tromper en affirmant que ce suremploi est caractéristique de l'univers d'"Europe", ou encore que les deux mots sont fortement associés dans le lexique du président (un indice 1 signifie simplement que L est supérieur ou égal à 0,9999999, la dernière décimale étant un 9, l'arrondi se fait à l'unité).

Le rapport entre l'effectif théorique (98,9) et l'effectif constaté (203) indique que le président avait une propension à associer "renforcer" avec "Europe" en moyenne deux fois plus forte qu'avec le reste de son vocabulaire. Cette proportion est même de 4 avec le verbe "élargir" ou l'adverbe "demain". Certaines associations vont de soi, notamment la présence (en dixième position) de "demain" qui est à mettre en relation avec la forte propension à placer au futur les verbes associés à "Europe". Il en va de même pour "modèle ("social européen")", "élargir", l'"euro" ou "monde" ("la place de l'Europe dans le monde").

D'autres associations paraissent plus mystérieuses comme la présence de "main", presque trois fois plus employée avec "Europe" que dans le reste du corpus. Pour comprendre cette association, il faut recourir aux concordances, en demandant à l'ordinateur les contextes associant "main" et "Europe". Voici, trois des premières lignes de cette concordance :

«… ma volonté que la France et la Hongrie marchent la **main** dans la **main** pour apporter leur contribution (…) à la construction de l'**Europe**, à la sécurité de notre continent » (Budapest, 16 janvier 1997)

«Il est évident que l'**Europe** et l'Amérique latine doivent marcher la **main** dans la **main**. (Assomption (Paraguay), 16 mars 1997).

« … une **Europe** assumant mieux ses responsabilités, la **main** dans la **main** avec une Amérique dynamique et ambitieuse » (Washington, 23 avril 1999).

Dans les phrases contenant "Europe", l'expression "main dans la main" est employée 38 fois – soit 76 des 94 occurrences de "main" dans cet univers -, 8 fois avec "la France et l'Allemagne", 6 fois avec "l'Europe et les Etats-Unis (ou l'Amérique)" et 5 fois avec "l'Europe et la Russie" ainsi que "l'Europe et l'Amérique latine" (souvent la France se trouve intercalée comme on le voit dans le premier exemple). On trouve ensuite l'expression "tendre la main" (6), "prendre en main" (5) dont 4 fois "son destin" et enfin "donner la main". Il s'agit de discours prononcés à l'étranger. Il est évident que, à chaque fois, le véritable sujet était le président Chirac parlant au nom de l'Europe ! Enfin comment ne pas se souvenir du "mano en la mano" de C. de Gaulle (Mexico, 16 mars 1964 ; voir à ce sujet l'article de E. Arnold dans ce même numéro) ?

Cet exemple permet de comprendre qu'il s'agit non pas de mots isolés mais d'associations de mots, soit dans une chaîne figée (comme "main dans la main", "modèle social", "construction de l'Europe", "politique commune"… "leadership de la France"), soit dans un rapport de substitution mutuelle ("paradigmes" ou famille de mots). Par exemple, puisque l'adjectif "pacifique" est associé à "Europe", on trouve "paix", en tête de liste, dans les substantifs C+ ; de même pour "renforcer", "renforcement" et l'adjectif "fort" ou encore "élargir", "élargissement" et l'adjectif "élargi"… Ces mécanismes d'association et de substitution permettent aussi de comprendre la présence de "alliance" ("atlantique") et de "Etats-Unis" parmi les vocables les plus fortement associés à "Europe" avec "Atlantique", "puissance" et "défense". Ils font également pressentir combien était capitale pour J. Chirac (et F. Mitterrand avant lui), la dimension militaire de la construction européenne et son insertion dans l'alliance avec les Etats-Unis.

La présence de "nord", "est" (nom masculin et non pas verbe), "centre", "continent" et "sud" (pour "Europe du nord, de l'est, etc." signale à la fois une conception géopolitique de la "construction européenne" mais surtout de ses "divisions" (autre substantif C+).

On note enfin que les deux pronoms les plus fortement associés à "Europe" sont le réfléchi "se" puis le pronom "nous" suivi de "elle-même" et de "on" (déjà évoqué). Dans des phrases comme "l'Europe se renforce, s'élargit, se construit, etc.", on peut parler d'un processus sans sujet (du moins sans sujet différent de l'objet). Nous allons montrer par la suite que plusieurs autres caractéristiques confirment cette impression d'un processus sur lequel le locuteur semble sans prise. Quant à "nous", dans les entretiens et conférences de presse sur l'Europe, il s'adressait quelquefois aux journalistes et signifiait "moi et vous à qui je parle". Pour le reste, la majorité de ces emplois désignaient le conseil européen (réunion des chefs d'Etat et de gouvernement de l'UE) et sinon un collectif à géométrie variable : tantôt "nous le gouvernement", tantôt "nous la France", voire "nous les Européens". En tous cas, ce "nous"

était une façon pour le président de s'effacer parmi d'autres. Les liaisons négatives en apportent confirmation.

5.3. Les principales liaisons négatives avec "Europe"

Quels mots sont significativement peu employés avec "Europe" ? Le tableau 10 donne les vocables les plus éloignés de l'univers et illustre la remarque selon laquelle il ne s'agit pas d'absence mais de présence significativement faible. Dans le tableau, le rang est inversé : le vocable le plus caractéristique est celui dont l'indice est le plus faible (quand cet indice est inférieur ou égal à 0,0000001, il est arrondi à zéro).

Tableau 10. Les dix vocables les plus sous-employés avec "Europe" dans les propos du président Chirac (classement en fonction de la force de la liaison)

Rang	Vocable	Catégorie gram.	Effectifs corpus	Effectifs univers	Effectif théorique	Proportion	Indice
3825	je	pro.	56 548	1878	2 816,4	0,033	0
3824	mon	dét.	7 961	194	476,2	0,024	0
3823	votre	dét.	9 947	288	595	0,029	0
3822	vous	pro.	23 869	791	1 188,8	0,033	0
3821	monsieur	n m	10 047	245	487,4	0,024	0
3820	public	adj.	3 159	50	173,6	0,016	0
3819	remercier	v.	2 624	31	135,3	0,012	0
3818	ce	dét.	38 029	1833	2 274,8	0,048	0
3817	loi	n. f.	1 774	13	86,1	0,007	0
3816	leur	dét.	11 525	465	689,4	0,04	0
(…)							
3809	Français	n. p.	4 228	142	261,7	0,034	0
3804	national	adj.	2 673	66	146,9	0,025	0
3803	français	adj.	4 982	161	273,8	0,032	0
3800	madame	n. f.	3 763	96	182,5	0,026	0

Le pronom "je" figure au premier rang des vocables les plus éloignés de l'Europe (il en manque près d'un millier par rapport à sa densité dans le reste du corpus). Quand il parlait de l'Europe, J. Chirac avait une propension moyenne à utiliser la première personne d'un tiers plus basse que pour les autres sujets. Autrement dit, quand il traitait ce sujet, le président s'absentait autant qu'il pouvait et assumait mal son discours. C'est la dimension principale de cet univers. Cette absence relative du locuteur est à mettre en relation avec la présence en C+ du "nous" et surtout du réfléchi qui aboutit ici à un procès sans sujet.

Le destinataire était également effacé autant que possible. C'est d'abord le "vous", spécialement "vous les Français" mais aussi "vous les journalistes" : dans ce dernier cas, le "vous" (et l'adjectif possessif votre) sont fondus dans le "nous". Le recul est considérable : pour

"Français" – 46% par rapport à la densité moyenne dans le reste du corpus. Comment mieux avouer que les Français n'étaient guère concernés par la "construction européenne" ?

Là encore, il faut considérer des familles de mots : l'adjectif "français" figure également dans les C- avec l'adjectif "national" qui, dans la bouche de J. Chirac, était un quasi-synonyme de l'adjectif "français".

La présence de "monsieur" – et donc de "madame" – dans les C- s'explique ainsi. Lorsque J. Chirac parlait d'une personne, spécialement quand c'était un responsable politique, il faisait précéder son nom de "monsieur" (ou "madame"). Donc la présence de ces vocables en C- signifie que le président nommait significativement peu de personnes physiques quand il parlait des problèmes européens. Il préférait utiliser le nom du pays, voire sa capitale. Mais cette absence relative des acteurs et leur remplacement par des entités – caractéristique bien connue du langage diplomatique – accentuait certainement le caractère abstrait des propos du président.

D'autres caractéristiques négatives peuvent être anecdotiques. Par exemple, "je vous remercie", formule rituelle pour mettre fin aux conférences de presse et aux entretiens. Elle forme une phrase entière, ce qui fait que le verbe "remercier" apparaît assez systématiquement en C- dans les univers des principaux vocables (sauf le pronom "je").

5.4. Le vocabulaire commun entre "Europe" et le reste du vocabulaire

Parmi les 3 825 vocables dont les fréquences sont supérieures aux différents seuils définis ci-dessus, 2 680 (soit 70 %) ne sont ni C+ ni C- ($0,05 < L < 0,95$). Certes, cette proportion est loin des 90% que laisserait attendre l'hypothèse sur laquelle se fonde le calcul mais cela montre que le raisonnement n'est pas sans fondement. Pour ces vocables, on accepte H_0 et on les considère comme communs au corpus et à l'univers. Le tableau 11 donne ceux dont les effectifs sont les plus élevés ($F_i > 1000$).

Etant donné la grande proximité entre les effectifs théoriques et constatés (effectif univers), il est certain que ces mots sont communs, c'est-à-dire que J. Chirac avait la même propension à les employer pour parler de l'Europe que pour les autres sujets.

Il est amusant de constater que l'adverbe "naturellement" vient en tête de liste. En effet, dans le vocabulaire du président, il s'agissait d'un synonyme de "évidemment". Il l'utilisait à chaque fois qu'il abordait un sujet difficile et dont la solution n'était pas évidente (ou "naturelle"). La liste révèle au moins deux autres formules très utilisées à tout propos : "matière à réflexion" et "réunion en cours".

Tableau 11. Les principaux vocables communs à l'univers de "Europe" et au reste du corpus (effectifs dans le corpus supérieurs à 1000)

Vocable	Catégorie gram.	Effectif corpus (F_i)	Effectif univers (F_{iu})	Effectif théorique	Proportion	Indice
naturellement	adv	4 714	255	255,4	0,054	0,5000
succès	n m	1 297	64	62,9	0,049	0,9494
matière	n f	1 219	61	59,1	0,05	0,9492
réflexion	n f	1 187	59	57,6	0,05	0,9477
côté	n m	1 141	52	55,3	0,046	0,0507
réunion	n f	1 107	51	53,7	0,046	0,0531
six	dét.	1 069	63	63,9	0,059	0,0514
cours	n m	1 051	48	51	0,046	0,0536
indispensable	adj	1 044	55	57,4	0,053	0,0524
diversité	n f	1 032	50	50,1	0,048	0,5000
attacher	v.	1 030	50	53,1	0,049	0,0524
langue	n f	1 029	51	49,9	0,05	0,9434

Plus bas dans la liste des vocables communs, on trouve "de Gaulle" avec un indice de 0,5 : sa densité dans l'univers est exactement la même que dans le corpus entier. Simple coïncidence ? Ou bien J. Chirac invoquait-il l'autorité du Général à tout propos et, spécialement, dans toutes les situations délicates ?

5.5. Phrases caractéristiques de l'Europe chez J. Chirac

Ces listes peuvent sembler assez abstraites, car les mots sont sortis de leur contexte. On propose donc un retour au texte, en demandant à l'ordinateur de trouver les phrases les plus caractéristiques de cet univers. Pour cela, le programme relit l'ensemble du corpus et classe les phrases, où figure le mot recherché, en fonction de la densité (absolue et relative) des C+ et C- y figurant. Ces phrases peuvent être considérées comme les citations canoniques que tout dictionnaire donne à l'appui de ses définitions. Ici, le choix n'est pas fait arbitrairement par le lexicographe mais il est effectué objectivement, de telle sorte que ces phrases sont certainement les plus caractéristiques du mot recherché.

Par exemple, la phrase ci-dessous prononcée le 10 octobre 1995, à Madrid dans une allocution adressée à la communauté française en Espagne.

« J'ajoute que l'Espagne et la France partagent une vision commune sur la plupart des problèmes, sur les problèmes européens d'abord et la construction de notre continent s'agissant de ses institutions, s'agissant des progrès à faire dans le domaine de l'économie et de la monnaie, s'agissant de l'élargissement, s'agissant aussi de nos relations avec nos voisins, nos voisins de l'Europe du nord, de l'Europe centrale et orientale, nos voisins Russes également, s'agissant enfin de nos relations avec l'ensemble méditerranéen qui auront l'occasion d'être mises particulièrement en exergue lors de la conférence de Barcelone, qui sera, je crois qu'on peut le dire, un événement historique, dans la mesure où il manifestera concrètement cette solidarité

que nous voulons voir s'établir et se renforcer entre l'ensemble des pays méditerranéens pour créer une zone de stabilité, de paix et de développement complémentaire de l'élargissement européen au nord et à l'est. »

Cette phrase compte 169 mots. Le solde des C+ et des C- est de 99. Elle éclaire bien la conception géopolitique de la construction européenne chez J. Chirac et montre également la propension à utiliser des phrases exceptionnellement longues quand le président abordait ce sujet. Le logiciel recherche également les phrases les plus courtes avec un score relatif remarquable qui se rapprochent un peu du slogan. Ci-dessous les trois premières (plus de 8 mots sur dix sont C+) :

«L'Europe est aujourd'hui un grand marché et une zone monétaire puissante. » (Allocution, Voeux aux Forces vives, 4 janvier 2005)

« La relation entre la France et l'Allemagne est une relation incontournable en Europe. » (Conférence de presse, Conseil européen, Hampton Court (Royaume Uni), 27 octobre 2005)

« Plus forte, plus rassemblée, plus dynamique, la France jouera en Europe un rôle moteur. » (Meeting de la campagne présidentielle, Lille, 18 avril 2002).

6. Dimensions stylistiques et grammaticales de l'univers de "Europe" comparé au reste du vocabulaire du président

L'univers de "Europe" chez J. Chirac présente des caractéristiques stylistiques et grammaticales singulières.

6.1. Les dimensions stylistiques de l'univers de "Europe"

En premier lieu, l'Europe donnait lieu à des phrases nettement plus longues que les phrases habituelles (tableau 12) Cette caractéristique est mesurée à l'aide de 4 valeurs centrales et un indice mesurant la dispersion des observations. Les deux dernières colonnes indiquent l'augmentation des indicateurs quand on passe du corpus entier aux univers de "Europe" ou de "Union Européenne".

La longueur de la phrase est mesurée au nombre de mots qu'elle contient et sans tenir compte des signes de ponctuation. Toute phrase supérieure à environ 15 mots est syntaxiquement complexe. Elle peut notamment comporter plusieurs propositions. Cette complexité syntaxique risque évidemment d'engendrer des problèmes de compréhension dans l'auditoire.

Tableau 12. Les longueurs de phrase (mots) valeurs centrales et déviation standard

	A Corpus entier	B Europe	C Union Européenne	B/A (%)	C/A (%)
Mode	14,0	19,0	21,0	+36	+50
Médiane	20,0	27,3	30,4	+37	+52
Moyenne	23,9	32,0	35,0	+34	+46
Médiale	29,9	38,1	40,0	+27	+34
Ecart-type	16,1	19,8	19,9	+23	+24
Variation relative	67,4	61,9	56,9	-8	-16

Les phrases sont rangées par longueurs croissantes dans des classes d'intervalles égaux (ici un mot). L'effectif de chaque classe est recensé et son poids relatif est calculé. Ce recensement fournit les informations suivantes :

Le *mode* (classe la plus peuplée) : longueur de phrase que le lecteur a le plus de chance de rencontrer : 14 mots dans le corpus entier. Il y avait donc, chez J. Chirac, une prédominance des phrases relativement courtes et syntaxiquement simples mais il en est ainsi dans la plupart des textes en français. Or cette valeur augmentait considérablement quand le Président parlait de l'Europe (+36%) et plus encore quand il évoquait l'Union Européenne (+50%). Dans ce dernier cas, les phrases les plus fréquentes du président ont 21 mots.

La *médiane* est la valeur de la variable pour l'individu du milieu ou individu "médian". Pour le corpus entier, la longueur médiane est de 20 mots, ou encore : la moitié des phrases ont une longueur inférieure ou égale à 20 mots et l'autre moitié une longueur supérieure à 20. Quand les phrases portent sur l'Europe, la médiane augmente de +37% et de + 52% pour "Union Européenne".

La *moyenne* : 23,9 mots n'est pas au milieu de la population comme on le croit souvent. En cas de distribution asymétrique (comme ici). Elle est tirée vers le haut par les individus les plus grands.

La *médiale* (ou "seconde médiane") est la valeur centrale la plus caractéristique d'une série statistique comme celle-ci : les phrases étant rangées par longueurs croissantes, c'est la longueur de la phrase qu'il faut atteindre pour couvrir la moitié du texte. Dans le corpus entier, elle est d'une trentaine de mots. Elle passe à 38 pour l'Europe et 40 pour l'Union Européenne. Autrement dit, quand J. Chirac parlait de l'Europe, son auditoire devait subir pendant la moitié du temps des phrases de longueur supérieure à 38 mots ce qui est beaucoup et excède certainement la capacité de compréhension de beaucoup, surtout parmi les non-spécialistes de ces questions.

La *dispersion* du phénomène ou déviation "standard" des valeurs de la variable autour de la moyenne. L'écart type de la longueur de l'ensemble des phrases du président Chirac est de 16,1

mots, mais de pratiquement 20 mots quand il parlait de l'Europe ou de l'Union Européenne. Cette mesure est complétée par le coefficient de variation relative : rapport de l'écart-type à la moyenne arithmétique (corpus entier : 67 %). Etant donné l'effectif considéré (171 201 phrases), si les valeurs de la variable "longueur de phrase" étaient distribuées normalement autour de la moyenne, ce coefficient serait inférieur à 4%.

Autrement dit, les observations sont très dispersées. Dans ce cas, la moyenne n'est pas représentative de la série et, en particulier, il n'est pas possible de considérer que cette moyenne se situe à peu près "au milieu" de la population. Dès que la dispersion relative approche les 50% de la moyenne, celle-ci est située dans la partie basse de l'étendue de la distribution qui est fortement asymétrique. En dernière colonne, le léger tassement du coefficient de variation relative indique que, quand il traitait de l'Europe, J. Chirac faisait des phrases de longueur un peu plus homogène que pour les autres sujets.

Tous les indices convergent pour indiquer un gonflement considérable de la phrase quand le président parlait de l'Europe. Il s'agit d'un automatisme : quand un sujet est particulièrement important, ou complexe, et qu'il occupe son esprit, tout locuteur a tendance à faire des phrases plus longues et plus compliquées qu'à l'ordinaire et à choisir un degré d'abstraction plus élevé, ce qui se traduit également par une prédominance du nom sur le verbe. En effet, l'allongement de la phrase est à mettre en relation avec un poids particulier de certaines catégories grammaticales.

6.2. Poids comparé des catégories grammaticales dans le corpus et l'univers de "Europe"

Lorsqu'il émet un énoncé, tout locuteur fait d'abord un choix fondamental : verbe ou nom ? Ce premier choix se traduit dans l'ensemble de la phrase. Une phrase centrée sur le verbe contiendra plus de pronoms, d'adverbes et de conjonctions de subordination. A l'inverse, le substantif est accompagné d'adjectifs, de déterminants et de prépositions. Dès lors, quand il parlait de l'Europe, J. Chirac avait-il la même propension à utiliser verbes et noms que dans le reste de ses propos ou privilégiait-il l'un ou l'autre de ces deux groupes ? Le tableau 13 donne la réponse (tableau détaillé en annexe 1).

Tableau 13. Poids des deux groupes de catégories grammaticales dans l'univers de "Europe" comparé au reste du corpus.

Catégories	A Corpus sans l'Europe (‰)	B Europe (‰)	A/B %	(B-A)/B %	Indice
Groupe du verbe	322,5	295,8	91,7	-8,3	0
Groupe du nom	675,9	702,4	103,9	+3,9	1

Groupe du verbe : verbes + pronoms + adverbes + conjonctions de subordination
Groupe du nom : noms propres – substantifs + adjectifs + déterminants + prépositions

Lorsqu'il parlait d'autres sujets, J. Chirac donnait en moyenne au groupe du verbe un poids de 322,5 ‰ (colonne A), mais quand l'Europe était le thème du propos, ce poids tombait à 295,8 ‰ (colonne B), soit un recul de 8,3%. La dernière colonne indique que l'on a moins de 1 chance sur 10 000 de se tromper en considérant que ce recul ne peut être dû au hasard. Naturellement, le poids du groupe nominal augmentait sensiblement dans les mêmes proportions. Là encore, l'indice en dernière colonne est de 1 (moins d'une chance sur 10 000 de se tromper en considérant qu'il y a "trop" d'éléments du groupe du nom dans l'univers d'"Europe".

Ce choix stylistique fondamental est nuancé grâce au détail des différentes catégories grammaticales (annexe 1).

Alors que les verbes étaient sous-employés, le futur était sur-employé de 2% et le participe présent de + 7,6%. La préférence pour le futur est attestée notamment par la présence en C+ de l'adverbe "demain" et elle est le complémentaire de la réticence manifeste à parler de l'Europe au conditionnel et au passé (imparfaits, passés simples, participes passés) voire au présent. Quant au participe présent, c'est la forme verbale la plus proche du substantif, celle qui implique le moins d'action ou un processus sans agent ni objet, caractéristique déjà relevée notamment à propos des pronoms réfléchis "se" et "elle-même".

L'importance des mots à majuscules s'explique d'abord par l'emploi des noms des pays – voire des capitales de ces pays – participant à l'Union, spécialement l'Allemagne.

Un déficit considérable en pronoms personnels (- 19,4%) : quand il est question d'Europe, le(s) sujet(s) s'absentent, plus particulièrement le "je".

L'excédent des déterminants ne provient que des articles. En particulier, lorsque J. Chirac parle de l'Europe, il emploie significativement peu de nombres. En effet, les dates et les chiffres ancrent le propos dans le temps, l'économie, la finance, la démographie… Leur faible présence accentue donc le côté abstrait du propos.

Les substantifs et les adjectifs sont également moins présents à propos de l'Europe par rapport au reste du discours chiraquien. Or ces deux catégories sont les principaux vecteurs de contenu avec les verbes.

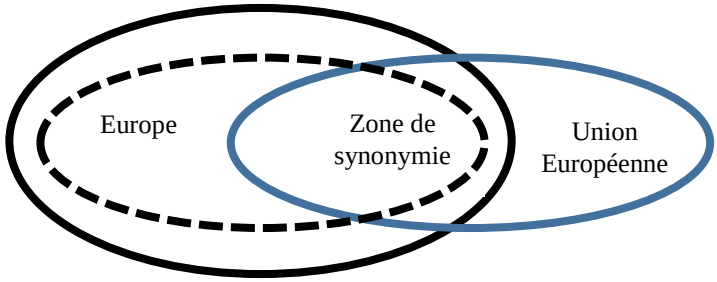
Au total, la densité des catégories grammaticales confirme la tendance de J. Chirac à produire des propos généraux et impersonnels quand il parlait de l'Europe. En est-il de même quand il désignait l'Europe à l'aide d'autres mots comme "Union Européenne" ?

7. Comparaison des univers

La méthode de comparaison des univers lexicaux a été présentée dans (Labbé, 1998). La principale difficulté réside dans les différences de taille : l'univers de "Europe" comporte 230 105 mots (et 6 285 vocables) contre 97 778 mots (et 4 050 vocables) pour celui de "Union Européenne". Autrement dit, l'un est plus du double de l'autre. Il n'est pas possible d'utiliser les fréquences relatives car un bon nombre de vocables présents dans l'univers de "Europe" ne franchissent pas les seuils qui leur permettraient de figurer également dans l'univers d'"Union Européenne".

On commence par réduire l'univers de "Europe", en divisant par 2,35 les effectifs des vocables caractéristiques et en ne retenant que ceux qui, dans cet univers réduit, ont une fréquence suffisante pour apparaître dans l'univers de "Union Européenne" (Graphique 2)

Graphique 2. Schéma de principe de la comparaison de deux univers lexicaux.



L'opération permet de distinguer plusieurs sous-populations : les vocables C+ et C- appartenant aux deux univers délimitent une "zone de synonymie" ; les vocables C+ et C– dans l'un seulement des deux univers sont propres à cet univers et définissent les sens particuliers qui différencient les deux.

Dans le vocabulaire de J. Chirac, les deux vocables n'étaient que partiellement synonymes. Lorsqu'il employait "Europe", il désignait une entité géographique et politique qui s'étendait de l'Atlantique à l'Oural, entité qui était divisée en zones géographiques (le nord, le sud, le centre, l'est) dominées par des problèmes politiques, diplomatiques, militaires et par les relations entre les principaux Etats à l'échelle mondiale (Etats-Unis, Russie, Chine). Secondairement, J. Chirac voyait dans l'Europe un vecteur pour réaffirmer la place de la France dans le monde et son "leadership". Le terme "Union Européenne" était préféré pour désigner le marché unique et lorsqu'il s'agissait de traiter les problèmes diplomatiques, financiers et économiques posés à cette entité.

8. Conclusions

Lorsqu'il parlait de l'Europe, J. Chirac s'impliquait faiblement dans ses propos. Ses phrases étaient nettement plus longues, plus impersonnelles et plus abstraites qu'à l'ordinaire. Enfin, il reconnaissait implicitement que les Français étaient peu intéressés par cette question à laquelle il consacrait pourtant une part considérable de son temps et une place primordiale dans sa communication. Certes, il en est ainsi depuis le général de Gaulle jusqu'à aujourd'hui : le président assume une relation politique personnelle avec les Français et se désengage plus ou moins des autres questions avec le "nous", voire l'impersonnel.

Cette caractéristique commune ne doit pas faire oublier qu'il existe de nombreuses singularités propres à chaque président. Des travaux ultérieurs, utilisant les mêmes méthodes appliquées à ces corpus présidentiels, permettront d'affiner les portraits lexicaux et stylistiques du général de Gaulle et de ses successeurs. En plaçant nos fichiers dans le domaine public, nous espérons susciter d'autres recherches en ce domaine.

Nous espérons également avoir montré combien la statistique appliquée au langage est un outil intéressant pour extraire de l'information dans une masse de textes équivalant ici à 31 forts volumes dont aucune technique "manuelle" ne peut venir à bout.

Elle peut également apporter une aide précieuse à la linguistique. Certes, le cadre intellectuel a été fixé il y a plus d'un siècle par Saussure : la langue est un système de systèmes dans lequel le sens d'un mot lui vient du réseau d'associations, d'exclusions ou de substitutions avec tous les autres éléments du système dans lequel il se trouve enserré. Mais le cerveau humain est mal outillé pour retrouver ces réseaux. Les mathématiques appliquées fournissent des formulations et des méthodes qui permettront de dépasser le stade intuitif, à condition de disposer de vastes corpus standardisés, étiquetés et indexés selon les méthodes évoquées au début de notre article.

Ces progrès possibles sont soumis à quelques exigences strictes : la standardisation et la rigueur des observations, la transparence des procédures et la reproductibilité des résultats.

Remerciements

Les corpus des interventions des présidents français (1958-2017) sont en ligne sur le site du Centre de Linguistique de Corpus (Université de Neuchâtel). La plupart des interventions des

présidents français depuis V. Giscard d'Estaing sont sur le site http://www.vie-publique.fr/. Les logiciels sont disponibles sur demande auprès des auteurs.

Les relecteurs anonymes ont permis d'améliorer ce texte.

Références

Les publications des auteurs sont consultables en ligne sur les "archives ouvertes du CNRS" (HAL) et sur researchgate

Arnold E. (2008). "Le sens des mots chez Tony Blair (people et Europe*).* In Heiden S. et Pincemin B. (Eds). *9es journées internationales d'analyse statistique des données textuelle*s. Lyon, Presses universitaires de Lyon, Vol. 1, p. 109 – 119.

Arnold E., Labbé C., Labbé D., Monière D. (2016). *Parler pour gouverner : Trois études sur le discours présidentiel français*. Grenoble, Laboratoire d'Informatique de Grenoble.

Chirac J. (2009). *Le temps présidentiel*. Paris, Nil.

Deligne A. (2013). *Le rôle de la France dans l'intégration européenne sous la présidence de Jacques Chirac (1995-2007).* Mémoire. École Nationale d'Administration publique.

Fabre C., Habert B., Labbé D. (1997). La polysémie dans la langue générale et les discours spécialisés. *Sémiotiques*. 13, décembre 1997, p. 15-30.

Hubert P., Labbé D. (1995). La structure du vocabulaire du général de Gaulle. *In Bolasco S., Lebart L. et Salem A. III Giornate internazionali di Analisi Statistica dei Dati Testuali.* Rome, Centro d'Informazione e Stampa Universitaria, 1995, II, p. 165-176.

Labbé C., Labbé D. (1994). *Que mesure la spécificité du vocabulaire ?* Grenoble, CERAT, décembre 1994 et juin 1997. Reproduit dans *Lexicometrica*, 3, 2001.

Labbé C., Labbé D. (2005). How to measure the meanings of words ? Amour in Corneille's work. *Language Resources Evaluation*. 39, p. 335-351.

Labbé C., Labbé D. (2010). La modalité verbale en français contemporain. Les hommes politiques et les autres. In Banks D. (éd.). *La modalité, le mode et le texte spécialisé*. Paris, L'Harmattan, 2013, p. 33-61.

Labbé C., Labbé D. (2017). *La répartition du vocabulaire*. Grenoble, Laboratoire d'informatique de Grenoble, septembre 2017.

Labbé D. (1990). *Normes de saisie et de dépouillement des textes politiques*. Grenoble, Cahiers du CERAT.

Labbé D. (1997). Le "nous" du général de Gaulle. *Quaderni di studi linguistici*. 4/5, p 331-354.

Labbé D. (1998). La France chez de Gaulle et Mitterrand. In Fiala P. et Lafon P. (dir). *Des mots en liberté. Mélanges Maurice Tournier*. Fontenay-aux-Roses, ENS Editions, 1998, p. 183-193.

Labbé D. (2010). *Le calcul du sens des mots. La lexicologie assistée par ordinateur*. Communication au séminaire Mathématiques et société. Neuchâtel, 3 novembre 2010.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1, p. 127-165.

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

Laurent A., Sauger N. (dir.) (2005). *Le référendum de ratification du Traité constitutionnel européen du 29 mai 2005*. Paris, Presses de sciences Po.

Lequesne C., Vaïsse M. (2012). *La politique étrangère de Jacques Chirac*. Paris, Riveneuve éditions.

Leselbaum J., Labbé D. (2002). Lexicographie assistée par ordinateur. Signification de "Banque" dans le vocabulaire économique. In Morin A. et Sébillot P. (Eds). *VIe Journées Internationales d'Analyse des Données Textuelles (Saint-Malo 13-15 mars 2002)*. Rennes, IRISA-INRIA, 2002, Vol. 2, p. 447-456.

Mandelbrot B. (1957). Étude de la loi d'Estoup et de Zipf. Fréquences des mots dans le discours. Apostel L. et al. *Logique, langage et théorie de l'information*. Paris, PUF, p. 22-53.

Mayaffre D. (2012). *Le discours présidentiel sous la Ve République*. Paris, Presses de Science po.

Monière D., Labbé C., Labbé D. (2008). "Les styles discursifs des premiers ministres québécois de Jean Lesage à Jean Charest". *Revue canadienne de science politique*. 41:1, p.43-69.

Monière D., Labbé D. (2012). Le vocabulaire caractéristique du Premier ministre du Québec J. Charest comparé à ses prédécesseurs. In Dister A., Longrée D., Purnelle G. (éds). *Proceedings of the 11th International Conference on Textual Data Statistical Analysis*. Liège, LASLA - SESLA, p.737-751.

Monière D., Labbé D. (2018). Le vocabulaire des campagnes électorales. In Iezzi D., Celardo L. et Misuraca M.. *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data. Roma*; UniversItalia, p. 522-540.

Mots. Les langages du politique (2010). Trente ans d'étude des langages du politique (1980-2010), 94.

Muller C. (1963). Le mot, unité de texte et unité de lexique en statistique lexicologique. Reproduit dans : *Langue française et linguistique quantitative*. Genève-Paris, Slatkine-Champion, 1979, p 125-143.

Muller C. (1977). *Principes et méthodes de statistique lexicale*. Paris, Hachette.

Péan P. (2007). *L'inconnu de l'Elysée*. Paris, Fayard.

Saussure F. de (1916). *Cours de linguistique générale*. Publié par Bally C. et Séchehaye A. avec Reidlinger A.. Réédition critique par Mauro T. de, Paris, Payot, 1993.

Savoy J. (2010). Lexical analysis of US political speeches. *Journal of Quantitative Linguistics*. 17, 2, p. 123-141.

Zipf G. K. (1935). *La psychobiologie du langage*. Paris : CEPL, 1974.

Annexe I.

Densités des catégories grammaticales dans l'univers de "Europe" comparées au reste du corpus

Catégories	A-B (Corpus-Univers) ‰	B (Univers) ‰	(B-A)/B %	Indice
Verbes	140.1	127.6	-9.0	0.0000
Futurs	5.3	5.4	+1.9	0.9904
Conditionnels	2.7	2.3	-16.1	0.0000
Présents	70.9	65.3	-7.9	0.0000
Imparfaits	5.0	4.9	-2.6	0.0084
Passés simples	0.4	0.3	-20.7	0.0000
Participes passés	20.4	16.2	-20.2	0.0000
Participes présents	2.7	2.9	+7.6	0.9969
Infinitifs	32.6	30.2	-7.6	0.0000
Mots à majuscule	24.2	61.0	+152.4	1.0000
Noms communs	196.5	167.7	-14.6	0.0000
Adjectifs	65.6	63.9	-2.6	0.0000
Adj. du part. passé	7.1	6.9	-2.2	0.0071
Pronoms	104.0	91.3	-12.2	0.0000
Pronoms personnels	52.9	42.8	-19.1	0.0000
Adverbes	61.9	59.4	-4.1	0.0000
Déterminants	191.9	204.4	+6.5	1.0000
Articles	139.2	157.7	+13.3	1.0000
Nombres	14.6	13.1	-10.4	0.0000
Adjectifs possessifs	19.4	17.5	-10.2	0.0000
Adjectifs démonstratifs	9.4	8.0	-15.2	0.0000
Adjectifs indéfinis	9.2	8.2	-11.0	0.0000
Prépositions	162.8	166.6	+2.3	1.0000
Conjonctions	51.5	56.4	+9.7	1.0000
Coordinations	35.0	38.8	+11.0	1.0000
Subordinations	16.5	17.6	+6.9	1.0000

Exemple de lecture : dans le corpus (auquel on a enlevé les phrases contenant Europe), les verbes ont une densité moyenne de 140,1 pour mille mots ; dans l'univers d'Europe, cette fréquence est de 127,6 ‰, soit - 9%. La dernière colonne indique que cet écart négatif a moins de 1 chance sur 10 000 d'être dû au hasard. C'est donc un fait stylistique significatif.

Annexe 2

Le vocabulaire caractéristique de l'Europe chez J. Chirac

1. Vocables significativement sur-employés dans l'univers (extraits)
(Seuil : 1, classement par catégories grammaticales et indices décroissants)

Noms propres :

- Peu fréquents : Oural, CEI, Niçois, Moratinos, Etchmiadzine, Asiatique, Balte, Chevtchenko,

- Fréquents : Traité de Rome, Renaissance, Dresde, Galileo, Oriental, Cadarache, Malte, Marché Commun

- Très fréquents : Etats-Unis, Européen, Amérique, Asie, Russie, Allemagne, MERCOSUR, Turquie, France, Maghreb, Berlin, Autriche, Méditerranée, Yalta, Sud-Est, Saint-Malo, Espagne, Balkan, Atlantique, Bangkok, Nice, OTAN, ASEM, Angleterre, Séoul, ITER, OSCE, Pologne, Chine, Hongrie, Amérique Latine, Singapour, Union Européenne, Caraïbe, Ukraine, Orient, Grande-Bretagne, Yougoslavie, Luxembourg, Roumanie, Japon, Caraïbes, Bundestag, Cannes, Serbie, Macédoine, Italie, Portugal, Strasbourg, Zagreb, Solana, Laeken, Irlande, Shoah

Verbes :

- Peu fréquents : arrimer, réimplanter, rivaliser, rééquilibrer,

- Fréquents: harmoniser, pencher, édifier, ancrer, léguer, apprêter, conjurer, bloquer, implanter,

- Très fréquents : renforcer, élargir, défendre, fonctionner, affirmer, enraciner, construire, doter, progresser, composer, avancer, devoir, bâtir, jouer, être, entendre, adhérer, promouvoir, relancer, contribuer, rejoindre, agir

Substantifs :

- Peu fréquents : leadership, coupure, métamorphose, hégémonie, trilinguisme, prêtre, prélat, fret, communisme, protéine, équipementier, théologie, natalité

- Fréquents : triangle, panne, sud-est, forteresse, chute, réunification, rideau, carrefour, unification, mémorandum, ancrage, fondateur, bloc, garante, sanctuaire, joug, dévaluation, nationalisme, destination, effacement, haut-représentant, repondération, retrouvaille, rivalité, épuration, subvention, concession, rééquilibrage, totalitarisme, attraction, avènement, complexe, synonyme, égal, édification, avant-garde, résurgence, libre-échange

- Très fréquents : main, modèle, alliance, euro, monde, nord, démocratie, intérêt, vocation, valeur, continent, pôle, élargissement, pays, scène, citoyen, puissance, paix, vision, moteur, construction, sud, est, ensemble, histoire, avenir, défense, union, sein, relation, identité, ambition, monnaie, lien, constitution, partenaire, stabilité, rôle, guerre, destin, politique, mur, frontière, partenariat, pas, croissance, prospérité, marché, progrès, voix, projet, présidence, division, place, fracture, traité, sommet, adhésion, idée, sécurité, réconciliation, nation, affirmation, fois, renforcement, géographie

Adjectifs :

- Peu fréquents : réunifié, apostolique, demeuré, ultralibéral, baltique, influent, asservi, paralysé, pacifié, inerte, intégriste

- Fréquents : continental, surmonté, technocratique, abstrait, interdit, euro-asiatique, forgé, rayonnant, germano-français, totalitaire, communiste, romain

- Très fréquents : pacifique, commun, fort, élargi, oriental, franco-allemand, économique, occidental, latin, uni, central, européen, multipolaire, réconcilié, grand, capable, atlantique, démocratique, méditerranéen, puissant, solidaire, prospère, transatlantique, monétaire, unique, ouvert, compétitif, social, fondateur, dynamique, asiatique, ambitieux, crédible, historique, exportateur, essentiel, stratégique, important

Pronoms :

- Peu fréquents : sien

- Fréquents : un, lui-même

- Très fréquents : se, nous

Adverbes :

- Peu fréquents : d'antan, culturellement

- Fréquents : après-demain, globalement, historiquement, contrairement, par-là, dorénavant, tantôt

- Très fréquents : demain, où, ensemble, plus, notamment, au-delà, enfin, clairement, aujourd'hui, seulement, longtemps, essentiellement, généralement, autour, davantage, oui, d'accord, aussi, considérablement

2. Vocables significativement sous-employés dans l'univers (extraits)
(Seuil : 1, classement par catégories grammaticales et par indices décroissants)

Noms propres :

- Peu fréquents : Québec, Raffarin, Serbe, Mali

- Fréquents : Jean-Pierre, Hugo, Guatemala, Hariri

- Très fréquents : Français, Nations Unies, Irak, Liban, ONU, Paris, G8, Israël, Algérie, Egypte, Polynésie, Corse, Libanais, Afghanistan, Lyon, Nouvelle-Calédonie, Moubarak, Syrie, Union Africaine, New York, Elysée, Beyrouth, Arménie, UNESCO, ONG, Niger, Evian, Georges, NEPAD, Sénégal, Africain, Jean, Pompidou, Irakien, Bush, Louvre, PME, Banque Mondiale, Réunion, Congo, Afrique du Sud, Johannesburg, Jacques, Clinton, Côte d'Ivoire, Michel, G7, Arafat

Verbes :

- Peu fréquents : vaincre, admirer, prolonger, préjuger, résister, envoyer

- Fréquents : remercier, falloir, saluer, adresser, prier, recevoir, demander, respecter, féliciter, répondre, reconnaître, prévoir, tenir, rencontrer, accompagner, réduire, venir, savoir, réserver, apprécier, renouveler, accueillir, agréer, exprimer

Substantifs :

- Peu fréquents : maître, être, scientifique, rigueur, chaleur, allusion, interdiction, mère, pandémie, transmission, expérimentation, académie, usager, inspection, médiation, accident, incitation, distance, tribunal, amour

- Fréquents : monsieur, loi, président, entreprise, société, madame, république, ministre, personne, vie, soin, maire, honneur, qualité, élu, travail, administration, plaisir, gouvernement, maladie, merci, résolution, mission, logement, eau, accès, accueil, métier, école, santé, secrétaire, action, prévention, estime, équipe, information, hommage, communauté, famille, activité, autorité, ami, malade, police, comité, lutte, essai, reconnaissance, assemblée, établissement, formation, émotion, risque, attention, financement, droit, armée, francophonie, compatriote, association, fonctionnaire, ministère, insécurité, humanité, égalité, délégation, parent, dévouement, quartier, hôpital

Adjectifs :

- Peu fréquents : reconnaissant, afghan, marocain, affectueux, représentatif, spécialisé, vietnamien, biologique

- Fréquents : public, cher, national, français, civil, personnel, chaleureux, général, heureux, professionnel, local, francophone, nucléaire, international, universel, sensible, privé, médical, familial, responsable, exceptionnel, républicain, africain, handicapé, rural, urbain, beau, vigilant, terroriste, sincère, législatif, haut, scolaire, nombreux, associatif, pauvre, cordial, âgé, administratif

Pronoms :

- Peu fréquents : moi, leur

- Fréquents : je, vous

Adverbes :

- Peu fréquents : volontiers

- Fréquents : ne, pas, encore, ici, très, beaucoup, particulièrement, là, souvent, auprès, hélas, autant, parfois, également, peut-être, ainsi, personnellement, trop, effectivement, pourtant, comment

3. Vocables non-caractéristiques ou communs au corpus et à l'univers (extraits)

Noms propres :

- Peu fréquents : PMA, KFOR, Auschwitz, Sierra Leone, Europe de l'Est, George, APD, Mohammed V, Kourou, José, Laval, Rhône, David, Koutchma, José Maria, EADS, Budapest, Banque Centrale Européenne, Albert, Voltaire, Thessalonique, Italien, Erika, Marshall, Riga, Hongrois, Hampton Court

- Fréquents : Roumain, Berlusconi, Javier, Polonais, Union Soviétique, Hollande, Charm El Cheikh, Bucarest, Rambouillet, Cardoso, Suédois, Dublin, République Tchèque

- Très fréquents : Lorraine, Maastricht, UEO, Paraguay, Sarajevo, Hu Jintao, Japonais, Britannique, TGV, Cologne, Caire, Estonie, Pays-Bas, Juif, Belgique, Chypre, Moldavie, Corée, Prague, Suisse, Russe, Allemand, Amsterdam, Kohl, Barcelone, Pékin, Royaume-Uni, Brésil

Verbes :

- Peu fréquents : prôner, hisser, brider, arracher, perdurer, propager, atteler, désintéresser, assimiler, déjeuner, pressentir, régresser, desservir, renverser, redéfinir, séjourner, finaliser, balayer, parachever, référer, réexaminer, endeuiller, ériger, refonder,

- Fréquents : trancher, comparer, paralyser, persister, méfier, perturber, heurter, refaire, afficher, combiner, innover, engendrer, déclarer, déboucher, verser, ménager, mêler, équiper, jouir, masquer, concrétiser, gouverner, redevenir, hésiter,

- Très fréquents : situer, substituer, rentrer, conférer, jeter, noter, souscrire, séparer, militer, correspondre, consolider, animer, figurer, signifier, apercevoir, aspirer, rejeter, percevoir, valoir, interroger, entretenir, chercher, libérer, confronter

Substantifs :

- Peu fréquents : espagnol, miracle, demeure, émancipation, ciment, desserte, finesse, fascination, rail, sucre, moratoire, gisement, visionnaire, maillon, nazisme, théorie, mortalité, onde, plénitude, nord-sud, renommée, supplément, vendredi, lord, important, larme, bien-fondé, dessous, cathédrale, mécontentement, maffia, rabais, carbone, ferveur, paralysie, chair, rancoeur, utopie, domination, foule, lecteur, exposé, intéressé, longueur, persécution, technicité, évêque, spatial, phare, maillage, aune, textile, modération, protectionnisme, prétention, cortège, angle, cursus, inscription, cinquantenaire, pourcentage, fusion, publication, concurrent, canton, proie, fenêtre, monopole, ruine, téléphonie

- Fréquents : substance, achèvement, coin, déchirement, prudence, réflexe, introduction, destinée, épidémie, refuge, excédent, avant-poste, voisinage, massacre, blessure, défenseur, ardeur, ferment, dictature, porte-parole, affinité, délocalisation, multiplication, clair, demandeur, successeur, primauté, quête, approvisionnement, petit, génocide, comparaison, pondération, prolongement, globalisation, illusion, chapitre, scepticisme, compte rendu, inflation, bailleur, séparation, professionnalisation, diplôme, camp, droite, haleine, alerte, donateur, quasi-totalité, pérennité, suppression, semestre, intuition, ailleurs, singularité, tournant, nationalité, aéronautique, rupture, débit, contour, golfe, renoncement, beauté, berceau, drapeau, signal, équivalent, monseigneur

- Très fréquents : occident, satellite, spécialiste, cérémonie, but, bord, spécificité, consultation, libéralisation, marge, automne, gros, courant, maximum, cycle, circulation, bout, intolérance, flux, prélèvement, interrogation, sacrifice, commissaire, idéologie, norme, bataille, lumière, phase, terroriste, mérite, contradiction, transition, hypothèse, table, proche, cohérence, blocage, disparition, docteur, référence, racine, totalité, habitude, vitalité, promesse, série, rayonnement, cap, thème, protocole, coup, critique, octobre, liaison, outil, intérieur, génie, exportateur, direction, échéance, souhait, affrontement, arrêt, théâtre, préparation, caractéristique, structure, absence, agriculteur, xénophobie

Adjectifs :

- Peu fréquents : danois, héroïque, artificiel, incessant, scandaleux, mondialisé, offensif, probable, végétal, impuissant, franco-japonais, semblable, conventionnel, bâti, pilote, fâcheux, recherché, dévasté, autre, salarié, obstiné, tissé, marquant, conjugué, protectionniste, restauré, motivé, infligé, inégalé, achevé, franco-italien, racial, paradoxal, rigide, inébranlable, manifeste, banal, arrivé, minimal

- Fréquents : préalable, passionné, visible, sévère, composé, ultrapériphérique, enthousiaste, producteur, perdu, maternel, pur, énorme, affectif, xénophobe, antisémite, net, libéré, anglais, éloigné, écrit, productif, tendu, grec,

bienvenu, inspiré, aéronautique, léger, présidé, loyal, triple, apporté, dépendant, turc, créatif, yougoslave, animé, suisse, propice, justifié

- Très fréquents : policier, mobile, formé, séculaire, prenant, roumain, incontournable, préoccupant, parcouru, informel, né, devenu, parfait, successif, renouvelé, potentiel, portugais, bas, musulman, industrialisé, renouvelable, planétaire, structurel, relatif, marqué, polonais, accru, protégé, maritime, énergétique, négatif, universitaire, insuffisant, brillant, unanime, contradictoire, mutuel, vivant, original

Adverbes :

Peu fréquents : à l'encontre, militairement, plain-pied, socialement, psychologiquement, dessus, réciproquement, spécialement, successivement, solidement, brillamment, financièrement, intimement, harmonieusement, singulièrement, par-dessus

Fréquents : conjointement, hautement, obligatoirement, facilement, massivement, nécessairement, relativement, infiniment, jadis, complètement, ardemment, cher

Très fréquents : finalement, activement, bref, véritablement, prochainement, fortement, suffisamment, largement, assez, certainement, réellement, normalement, évidemment, pleinement, précisément, ensuite, combien, naturellement, définitivement