HAL
open science

Existe-t-il un langage propre à la politique ?

Cyril Labbé, Dominique Labbé

To cite this version:

Cyril Labbé, Dominique Labbé. Existe-t-il un langage propre à la politique ?. Nouvelles Journées de l'ERLA n° 12, Nov 2011, Brest, France. halshs-00644646

HAL Id: halshs-00644646

https://shs.hal.science/halshs-00644646v1

Submitted on 24 Nov 2011

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.

Université de Bretagne occidentale
Faculté des Lettres et Sciences sociales
Equipe de Recherche en Linguistique Appliquée

Nouvelles Journées de l'ERLA n° 12

Aspects linguistiques du texte politique
(Brest 18-19 novembre 2011)

Existe-t-il un langage propre à la politique ?

Cyril Labbé
Laboratoire d'Informatique de Grenoble - Université Joseph Fourier
(cyril.labbe@imag.fr)

Dominique Labbé
Laboratoire PACTE (CNRS - Institut d'Etudes Politiques de Grenoble)
(dominique.labbe@iep.grenoble.fr)

Résumé
Après avoir présenté le corpus du discours politique français et la bibliothèque du français moderne, on recherche les caractéristiques singulières du discours politique. Le cas des présidents de la Ve République depuis 1958 est particulièrement étudié. Les vocables usuels ne sont pas propres à la politique mais leurs fréquences d'emploi et leurs combinaisons dessinent plusieurs univers singuliers. Le discours politique est assez dépersonnalisé avec une tonalité nettement pédagogique. Il est orienté vers le présent et le futur. Il privilégie les généralités et les abstractions plutôt que l'action. L'expérience montre l'utilité de la statistique et des vastes corpus de référence pour la linguistique appliquée.

Abstract
Presentation of the corpus of French political discourse which is a section in the library of modern French. The focus is placed on the presidents of the Fifth Republic since 1958. The usual words are not unique to politics, but their frequencies and their combinations create some characteristic lexical universes. Political discourse is rather impersonal with a pedagogical tone. It is oriented towards the present and the future. It focuses on generalities and abstractions rather than on the action. This experiment shows the usefulness, for applied linguistics, of statistics and of large corpora of reference.

Mots clefs : France – Québec - Canada – discours politique – président de la république – vocabulaire – lexique - statistique lexicale.

Cette année le choix des organisateurs des Journées de l'ERLA amène à se poser une question préalable : *Existe-t-il un langage propre à la politique* ? C'est en examinant cette première question que l'on pourra mettre en évidence les éventuels aspects linguistiques singuliers du texte politique.

Certes, il existe des professionnels de la politique, des lieux, des institutions propres à cette activité. Logiquement, ils devraient avoir développé un langage particulier. Mais peut-on s'en tenir à cette simple intuition ? On pourrait tout aussi bien soutenir que, gouvernant au nom de la population française toute entière et condamnés à solliciter régulièrement les suffrages de leurs électeurs, les hommes politiques sont, au contraire, condamnés à parler "comme tout le monde".

Pour apporter une réponse sûre, il faudrait disposer d'une collection de textes émis par un grand nombre d'hommes politiques et d'un étalon de comparaison : un vaste échantillon représentatif du français "général", c'est-à-dire commun à toute la population.

Cette communication présente une expérience qui simule cette comparaison, apporte quelques éléments en faveur de l'existence d'un discours propre à la politique et met en lumière ses singularités.

I. Des textes aux corpus

La seule enquête scientifique sur l'usage du français date de plus d'un demi-siècle. Elle a été pilotée par G. Gougenheim (1900-1972). Au début des années 1950, lui et sa petite équipe ont enregistré des locuteurs de tous les milieux, à propos de leur vie quotidienne, de leur travail de leurs loisirs…, puis ils ont saisi ces enregistrements et en ont réalisé un traitement statistique qui a abouti à une grammaire élémentaire du français et à un vocabulaire fondamental contenant les 3 500 mots les plus utilisés du français accompagnés des phrases canoniques (Gougenheim, 1956 et 1958). Avec les moyens de l'époque, c'était un travail remarquable, malheureusement resté sans suite, de telle sorte que le français est la seule grande langue de culture pour laquelle on ne dispose d'aucun corpus représentatif des usages, constitué sur le modèle du British national corpus.

Depuis 30 ans, nous avons constitué une "bibliothèque électronique du français contemporain" (annexe 1). Cette bibliothèque – qui compte actuellement 23 millions de mots - a été constituée au gré des collaborations avec des littéraires, des politistes, des spécialistes des médias (pour la presse), des sociologues et des psychologues (pour le français oral) et des informaticiens (pour l'écriture des programmes). Elle n'a pas les dimensions ni la représentativité d'un véritable corpus de langue mais, grâce à sa diversité et à son étendue, elle permettra tout de même d'apporter une première réponse à la question de départ.

On utilisera d'abord la section "discours politique" de cette bibliothèque soit 3 130 discours prononcés par plus d'une centaine d'hommes politiques du XXe siècle, non seulement Français mais aussi Québécois et Canadiens. Cette section comprend au total plus de 8,8 millions de mots étiquetés. Elle sera comparée aux autres sections de la bibliothèque, notamment la littérature et le français oral

1.1 Traitements préalables

Avant son entrée dans la bibliothèque, chaque texte subit une série de traitements (décrits dans Labbé 1990b).

En premier lieu, des balises sont ajoutées pour indiquer les sources du texte (auteur, titre, lieu, date… comme dans un catalogue), pour délimiter les séquences, isoler le texte du "para-texte" – pour les textes politiques : début et fin des propos des orateurs, interruptions, questions et réponses pour les interviews, etc. Ce travail est indispensable pour ne pas confondre les propos de l'homme politique – qui font seuls l'objet des analyses présentées ici - avec ceux des autres participants qu'il est pourtant essentiel de conserver pour la bonne compréhension du texte.

En second lieu, l'orthographe a été soigneusement corrigée et les graphies multiples ont été standardisées. Par exemple, "événement" et "évènement" ne sont qu'un même mot écrit de deux manières différentes, de même que "puis" et "peux", etc. Cela concerne particulièrement les sigles, les abréviations, les noms propres, les chiffres et les dates dont la transcription est d'une infinie variété… Prenons un exemple d'actualité. Dans les textes politiques depuis les années 1970, on rencontre les personnes suivantes : Khadafi, Kadhafi, Gadhafi, Gaddafi, al-Gaddafi. Il y a probablement d'autres manières encore de transcrire ce nom. Pour Mao, dans les textes qui ont alimenté notre modeste base, il y a 8 graphies différentes. Si l'usager de la bibliothèque électronique recherche tous les textes qui parlent de Kadhafi ou de Mao, que doit-il faire ? La réponse est double. Premièrement, dans l'index – qui sert de porte d'entrée à la bibliothèque - toutes ces graphies sont ramenées à une seule : la graphie standard (ou acceptée par la majorité des lexicographes). Deuxièmement, si l'usager demande une graphie non conventionnelle, il faut lui suggérer la conventionnelle…

Enfin, pour toutes les statistiques, le principe est le même qu'en démocratie : chaque individu compte pour un dans le vocabulaire ! Sinon, la base serait encombrée de fantômes et elle serait impropre à la recherche lexicographique, linguistique ou statistique.

Ces tâches sont partiellement effectuées par des automates, mais les interventions manuelles sont nombreuses et suivent des règles précises. Il en est de même pour la seconde opération : l'étiquetage des mots.

1.2 Etiquetage des mots

En français, beaucoup de mots sont ambigus. Par exemple, puisque l'on s'intéresse au discours politique, voici quelques exemples d'ambiguïtés gênantes avec le nombre de fois que le mot apparaît dans la bibliothèque.

est (verbe être) : 123 252
est (substantif masculin singulier) : 624

devoir (verbe infinitif) : 335
devoir (substantif masculin singulier) : 1905

pouvoir (verbe à l'infinitif) : 2005
pouvoir (substantif masculin singulier) : 2595

savoir (verbe infinitif) : 1488
savoir (substantif masculin singulier) : 718

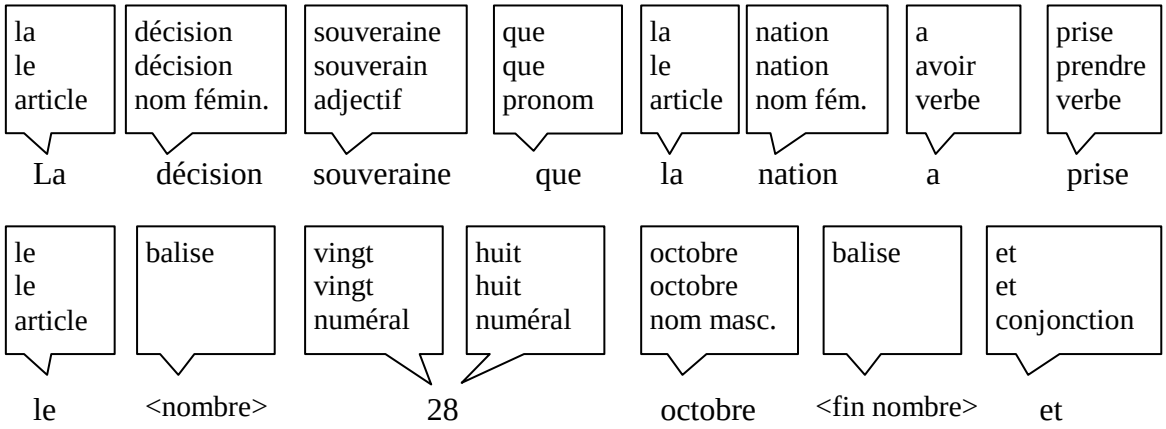
suis (verbe être) : 8057
suis (verbe suivre) : 50

Faut-il demander au chercheur en science politique de lire des milliers de lignes de "concordances" en écartant manuellement les verbes *pouvoir* à l'infinitif – qui ne l'intéressent pas - pour retrouver le substantif homographe qui est au cœur de ses études ? De même pour le *devoir* ou le *savoir*…

Or, dans tout texte en langue française, en moyenne un tiers des mots sont des homographes et ce sont souvent les mots les plus fréquents. A cela, il faut encore ajouter que 15 % des mots en moyenne sont des verbes dont les flexions (et les homographies) sont si nombreuses qu'il est impossible de les retrouver dans les bases de textes en ligne, ce qui les rend peu utiles pour les recherches lexicologiques.

La solution est la suivante. Le texte est découpé en autant d'emplacements (en anglais "tokens") qu'il y a de mots et chacun de ces emplacements est doté d'une étiquette indiquant l'entrée sous laquelle se trouve le mot dans un dictionnaire de langue (en anglais "type"). On désigne cette opération sous le nom de "lemmatisation" mais "étiquetage" serait préférable. Le tableau 1 ci-dessous donne un exemple d'étiquetage.

Tableau 1. Etiquetage des premiers mots de l'allocution radiotélévisée du général de Gaulle du 7 novembre 1962.



Dans ces étiquettes, la première position est occupée par la graphie standard, la seconde par l'entrée du dictionnaire (par exemple, l'infinitif du verbe) et la troisième par la **catégorie grammaticale** qui figure après l'entrée de dictionnaire. Le **vocable** est l'association de l'entrée de dictionnaire et de la catégorie grammaticale. La nomenclature est celle des dictionnaires de langue. Elle respecte les caractéristiques exigées d'une nomenclature statistique : univoque, exhaustive et sans double compte. La lemmatisation est en outre réversible : on peut retrouver le texte original à partir des étiquettes.

L'étiquetage est confié à des automates, mais ceux-ci ne peuvent pas tout faire, du moins si l'on recherche une lemmatisation sans faute, ce qui est notre objectif. Naturellement, ces opérations préalables sont longues et doivent être menées avec soin. Cela explique les dimensions relativement réduites de notre bibliothèque. En contrepartie, grâce à ces étiquettes, on peut établir le vocabulaire d'un auteur, d'une profession, d'un genre. Commençons par un exemple.

II. Un exemple, les vocables les plus employés par les présidents de la Ve République

C'est un postulat bien ancré que chaque homme, chaque profession possède son vocabulaire et qu'il suffit d'observer les mots les plus utilisés par ce ou ces locuteurs pour détecter leur vocabulaire. Le tableau 2 ci-dessous présente les vocables les plus employés par N. Sarkozy (dans les discours des 4 premières années de sa présidence), comparé à ses prédécesseurs. C'est ce que l'on nomme un index "hiérarchique".

Tableau 2. Les vocables les plus employés chez N. Sarkozy comparé aux autres présidents de la Ve République.

	Sarkozy			Autres Présidents		
Rang	Lemme et catégorie grammaticale	Effectifs	Fréquence (‰)	Lemme et catégorie grammaticale	Effectifs	Fréquence (‰)
1	le (det)	255 592	111.4	le (det)	99 775	105.0
2	de (pré)	175 903	76.7	de (pré)	69 580	73.2
3	**être (v)**	76 418	33.3	**être (v)**	31 289	32.9
4	à (pré)	56 160	24.5	à (pré)	25 080	26.4
5	**avoir (v)**	52 358	22.8	et (cj)	22 677	23.9
6	et (cj)	47 540	20.7	**avoir (v)**	20 460	21.5
7	un (det)	45 151	19.7	**je (pro)**	17 481	18.4
8	**je (pro)**	43 268	18.9	que (cj)	17 361	18.3
9	que (cj)	38 510	16.8	un (det)	17 167	18.1
10	ce (pro)	33 870	14.8	**il (pro)**	14 666	15.4
11	ne (adv)	33 075	14.4	ce (pro)	13 518	14.2
12	**il (pro)**	27 609	12.0	ne (adv)	12 592	13.3
13	pas (adv)	26 312	11.5	qui (pro)	12 558	13.2
14	qui (pro)	26 110	11.4	pas (adv)	9 726	10.2
15	pour (pré)	24 010	10.5	en (pré)	9 603	10.1
16	**nous (pro)**	21 269	9.3	pour (pré)	7 939	8.4
17	en (pré)	20 859	9.1	**nous (pro)**	7 818	8.2
18	**on (pro)**	19 314	8.4	ce (det)	7 455	7.8
19	ce (det)	17 019	7.4	dans (pré)	7 447	7.8
20	**vous (pro)**	16 351	7.1	se (pro)	7 273	7.7
21	dans (pré)	16 215	7.1	que (pro)	7 167	7.5
22	que (pro)	16 044	7.0	le (pro)	6 283	6.6
23	**faire (v)**	13 845	6.0	**on (pro)**	6 005	6.3
24	se (pro)	13 609	5.9	**faire (v)**	5 139	5.4
25	le (pro)	12 342	5.4	mais (cj)	4 946	5.2
26	**dire (v)**	11 144	4.9	y (pro)	4 716	5.0
27	y (pro)	11 084	4.8	**vous (pro)**	4 702	5.0
28	plus (adv)	10 907	4.8	**dire (v)**	4 532	4.8
Total		1 161 888	506.6		474 955	499.8

Comme indiqué dans l'annexe 1, pour N. Sarkozy, le corpus comporte toutes les interventions mises en ligne sur le site de la présidence, alors que pour ses prédécesseurs, il n'y a que les interventions radio-télévisées et les conférences de presse. Ces corpus seront complétés, dans la mesure où des archives existent et sont accessibles.

Cette réserve admise, le tableau appelle trois remarques.

Premièrement, pour rendre les deux colonnes comparables, les **effectifs** – chiffres absolus – sont transformés en **fréquences** (chiffres relatifs). Ces chiffres relatifs sont exprimés en "pour mille mots". Même ainsi, les proportions deviennent rapidement assez petites quand on descend dans la hiérarchie. Chez Sarkozy, comme chez les autres, on passe au-dessous de 1% après le 15e vocable. De plus, la surface du texte est très inégalement répartie. La dernière ligne du tableau indique que les 28 vocables les plus fréquents couvrent, à eux seuls, la moitié de cette surface. Et les 19 000 restants se partagent donc l'autre moitié mais ce sont eux qui véhiculent la plupart des informations !

Il en est toujours ainsi de telle sorte que, du point de vue statistique, tout texte produit en langue naturelle peut se définir comme une population constituée d'un grand nombre d'individus rares et d'effectifs très inégaux.

La deuxième remarque concerne la présence massive, en tête de liste, des "mots outils" (articles, prépositions, adverbes, pronoms). Dans les mots les plus employés, aucun substantif ni adjectif. En caractères gras, les cinq pronoms (*je*, *il*, *nous*, *on* et *vous*) et les quatre verbes (*être*, *avoir*, *faire* et *dire*). A vrai dire, seule la présence de *dire* est significative du discours politique car, dans tout corpus de textes français, les trois verbes les plus employés sont toujours *être*, *avoir* et *faire*, dans cet ordre.

Le premier substantif, *France* – que les présidents emploient 9 397 fois soit 4,1 fois pour mille mots – figure au 35e rang. Le second, *pays* (8 479 occurrences), représente 2,6‰ mots et vient au 45e rang, etc. Si l'on admet le postulat implicite de la plupart des études lexicologiques, selon lequel les "mots outils" sont des formes vides et que l'essentiel de la communication se fait à l'aide des verbes, des noms et des adjectifs, le tableau… signifie que, dans la vaste population des mots, on s'intéresse aux plus rares ! Ce qui pose un sérieux problème au statisticien. Si l'on considère les 91 985 vocables différents présents dans les textes de la bibliothèque, on constate qu'il y en a 54 896 – soit 6 sur 10 – qui sont utilisés moins de 5 fois. Pour ceux-là, la seule chose que l'on peut dire est qu'ils sont rares ! Il est vrai que, parmi eux, la majorité sont des mots à majuscule dont la nature particulière sera discutée plus loin. Il n'en reste pas moins que les opérations statistiques usuelles ne pourront porter que sur une petite partie du vocabulaire total…

Troisième remarque, il n'y a pratiquement aucune différence entre la partie gauche et la partie droite du tableau 2 surtout si l'on s'intéresse au rang qui suggère un ordre de préférence implicite et montre que Sarkozy semble partager cet ordre avec ses prédécesseurs. Cette moyenne cache d'ailleurs quelques différences, notamment chez de Gaulle qui employait plus la première personne du pluriel (*nous* : 9,0 ‰) que celle du singulier (*je* : 8,1‰).

Cependant, il faut descendre plus bas dans les listes pour voir apparaître des différences plus sensibles, notamment entre les hommes politiques et le reste de la population.

III. Le vocabulaire des hommes politiques est-il singulier ?

Pour répondre à cette question, on utilise un raisonnement statistique simple. Pour épargner les auditeurs à qui il a été présenté l'année dernière (Labbé & Labbé 2010), nous renvoyons à nos travaux disponibles en ligne : le calcul du vocabulaire caractéristique est présenté dans Labbé & Labbé 1994 et Monière, Labbé & Labbé 2004.

Le raisonnement consiste à considérer la bibliothèque dans son ensemble comme une "urne" et le corpus "discours politique" comme un "échantillon" extrait de cette urne. On compare les effectifs de chacun des vocables dans l'échantillon et dans l'urne. Pour chaque vocable, le test statistique permet de choisir entre deux hypothèses contradictoires : l'utilisation qu'en font les hommes politiques ne s'écarte pas significativement de l'usage général ou, au contraire, les hommes politiques s'écartent significativement (en plus ou en moins) de cet usage général. Pour cette étude, le seuil à partir duquel on accepte l'hypothèse d'un écart significatif a été fixé à 1% (de risque d'erreur), ce qui est sévère. Malgré tout, on obtient des listes très longues (plus d'une dizaine de pages dans le cas présent). L'annexe 2 ne reproduit que les vocables les plus caractéristiques.

Ces listes appellent quelques remarques.

Première remarque : les mots forment des groupes (**paradigmes**). Personne ne sera surpris que le substantif *politique* (nom féminin) soit sur-employé par les politiques. Logiquement, l'adjectif *politique* est également sur-utilisé. On a beaucoup d'autres couples de ce genre comme *économie-économique*, *nation-national*, etc. De même puisque les pronoms démonstratifs sont sur-employés par les hommes politiques (*ce, celui, cela*) alors aussi les articles *ce* et *cet* le sont aussi. Si *nous* est un pronom privilégié, alors ce sera aussi le cas de *nous-même(s)* – et de *nous autres* au Québec – du pronom *nôtre* ainsi que de l'adjectif possessif *notre*. Puisque *France* est le nom propre le plus employé, il est logique de trouver l'adjectif *français* en bonne place dans les vocables sur-employés, de même pour *Québec-québécois* et *Canada-canadien*. Il en est de même pour les verbes. Par exemple, le substantif *travail* vient toujours avec le verbe *travailler*, l'adjectif *possible* avec le verbe *pouvoir*, etc. Il faut donc étudier des paradigmes et non des unités de vocabulaire isolées.

Deuxièmement, ces listes sont classées par catégories grammaticales car chacune de ces catégories remplit une fonction particulière. Par exemple, les mots à majuscule initiale (patronymes, toponymes, sigles…) assurent l'ancrage spatial et social du discours. Tout naturellement, le nom du pays est le mot le plus employé et le plus caractéristique du discours politique dans chacun des trois pays (*France*, *Québec*, *Canada*) et les plus sous-employés sont les personnages de roman, les sigles du vocabulaire économique...

Les substantifs et les adjectifs présentent les principaux thèmes : institutions (*gouvernement*, *président*, *ministre*) la *politique* de…, le (ou les) *pays*, et la préoccupation principale depuis plus de 30 : l'*emploi*, le *travail*, l'*économie*, le *développement*…

L'adjectif prolonge le nom. A ce sujet, il est intéressant de noter que, dans toute grande collection de textes en français - entretiens, articles de presse ou œuvres littéraires - les deux adjectifs les plus employés sont "grand" et "petit" puis viennent d'autres couples comme "jeune-vieux", "beau-laid", etc. Une exception : dans le discours politique, on ne trouve que le terme marqué positivement (*grand* mais pas *petit* ; *jeune* ou *nouveau* mais pas *vieux*, etc.).

Troisièmement, les écarts les plus considérables sont observés dans la catégorie des pronoms, spécialement les pronoms personnels et démonstratifs (tableau 3 et 4).

Tableau 3. Densité des pronoms personnels dans le reste de la bibliothèque et dans le discours politique

Pronoms	A ‰ (Autres corpus)	B ‰ (Discours politique)	(B – A) / A %
je	22.01	13.05	-40.7
moi	2.52	0.76	-69.8
mien	0.12	0.04	-63.7
il	20.25	11.23	-44.5
lui	3.04	0.70	-77.1
se	8.93	6.14	-31.2
sien	0.09	0.05	-45.2
on	7.89	5.98	-24.2
nous	3.19	10.35	+224.6
nôtre	0.04	0.13	+270.8
vous	7.63	5.33	-30.1
vôtre	0.07	0.06	-21.0
ils	4.01	2.89	-28.0
Moyenne pronoms pers.	83,3	57,2	-31.4

Le tableau se lit de la manière suivante : dans les sections de la bibliothèque autres que le discours politique, on rencontre 22 pronoms de la première personne du singulier, pour 1000 mots, contre 13 ‰ dans les discours politique, soit 40,7 % de moins. La dernière ligne indique que, en moyenne, les hommes politiques utilisent environ 31 % de pronoms personnels de moins que les autres (c'est-à-dire 7 quand les autres en utilisent 10). Cette moyenne cache des mouvements contradictoires. Toutes les personnes sont sous-employées sauf la première du pluriel (*nous*) dont la fréquence est 3,25 fois plus élevée dans le discours politique que dans les autres corpus. Quels que soient les fonctions et les hommes, l'usage privilégié du *nous* semble être un invariant du discours politique. Ce *nous* a une infinité de nuances (Labbé 1997) – allant du simple pluriel de majesté, fréquent chez les présidents de la Ve, à l'expression de la nation, en passant par l'équipe gouvernementale – sens le plus fréquent chez les premiers ministres d'Amérique du nord.

En définitive, la principale caractéristique du discours politique serait la prétention à "incarner" des collectifs et, en premier lieu, à prêter une voix à la communauté nationale. Si, depuis G. Pompidou, les présidents français utilisent un peu plus *je* que *nous*, chez eux aussi, la densité d'emploi du *je* est, en moyenne, nettement inférieure à ce que l'on rencontre dans les autres corpus et, comme tous les autres hommes politiques, ils évitent autant que possible *moi* et *le mien*.

Dans les différents types de pronoms, il y a une seule exception au sous-emploi de cette catégorie par le discours politique : les pronoms démonstratifs (tableau 4).

Tableau 4. Densité des pronoms démonstratifs dans le reste de la bibliothèque et dans le discours politique

Pronoms	A ‰ (Autres corpus)	B ‰ (Discours politique)	(B – A) / A %
ça	3.7	0.9	-76.8
ce	11.1	11.0	-1.1
ceci	0.1	0.1	+44.7
cela	0.8	2.2	+168.6
celui	1.2	2.2	+81.9
Moyenne pronoms démonst.	13,3	15,6	+17.4

Ce tableau se lit comme le précédent. Le « ça » appartient au registre familier, de la conversation, qui ne convient pas dans le discours public soutenu. En revanche, les autres pronoms démonstratifs sont les marques d'un discours à visée pédagogique : ils servent au maître à placer son propos dans l'ordre des choses (*il faut… ; c'est comme ça*). L'homme politique se conçoit donc un peu comme un professeur qui tente d'expliquer le mieux possible à son auditoire ce que celui-ci doit savoir sur les grandes questions qui se posent à la nation.

On commence donc à voir apparaître les singularités du discours politique. Mais il faut bien avoir conscience que le sens des vocables provient de leur appartenance à des paradigmes (déjà évoqués) mais surtout de leur combinaisons avec d'autres vocables (**syntagmes**) pour former des énoncés. C'est donc les combinaisons de mots qui pourraient donner la réponse à la question de savoir si le vocabulaire politique est réellement singulier.

IV. Les combinaisons de mots préférées des politiques.

Les lexicographes désignent par le mot "collocations" les combinaisons de mots qui reviennent à l'identique à plusieurs reprises (Blumenthal & Al. 2005). Mais il s'agit de mots et non de vocables. Ainsi : *j'ai voulu dire, je veux dire* ou *cela ne veut pas dire* sont trois collocations différentes. L'étiquetage, décrit au début de cette communication, permet de reconnaître dans ces trois collocations la même combinaison des verbes *vouloir* et *dire*. La notion de **syntagme répété** désigne cette opération consistant à envisager toutes les constructions possibles et à regrouper toutes les flexions, notamment des verbes (Pibarot et Al. 1988 ; Labbé & Labbé 2010). Par exemple, « (je ne avoir) *vouloir* (pas, plus, bien) *dire* », etc. Au-delà de ces combinaisons, l'association des mots entre eux dessine des **univers lexicaux** (Hubert & Labbé 1995 ; Labbé & Labbé 2005).

Le tableau 5 présente les principaux syntagmes répétés des présidents français comparés à une autre section de la bibliothèque (la littérature). Précisons bien que les auxiliaires *être* ou *avoir* précédant un participe passé ne sont pas présents dans ce tableau. Par exemple, des constructions comme "il peut avoir dit", "il n'a pas pu dire" sont rattachées au syntagme "pouvoir dire".

Tableau 5. Les principaux syntagmes répétés des présidents français comparés à ceux de la littérature française

Rang	Présidents	Fréquence (pour 10 000 mots)	Littérature française	Fréquence (pour 10 000 mots)
1	vouloir dire	13,1	pouvoir être	4,6
2	devoir être	13,0	être vrai	4,5
3	avoir besoin	11,2	être grand	4,0
4	pouvoir être	10,3	avoir besoin	4,0
5	madame monsieur	8,6	être bon	4,0
6	France être	6,3	devoir être	3,6
7	président république	6,0	pouvoir faire	3,5
8	pouvoir dire	5,9	être beau	3,2
9	être grand	5,3	être seul	3,0
10	être important	5,3	avoir peur	2,8
11	pouvoir faire	5,2	jeune homme	2,8
12	être question	5,0	avoir air	2,8
13	être vrai	5,3	être heureux	2,7
14	chef état	4,8	être homme	2,7
15	monsieur président	4,7	être sûr	2,5
16	avoir droit	4,4	vouloir dire	2,4
17	être bon	3,9	avoir raison	2,3
18	être heureux	4,3	aller faire	2,0
19	être pays	4,0	vouloir faire	2,0
20	pouvoir avoir	3,9	pouvoir dire	1,9

Le tableau se lit de la manière suivante : chez les présidents de la république, le syntagme le plus souvent répété est *vouloir dire*. En moyenne, les présidents en ont utilisé 13 par tranches de 10 000 mots. Dans la littérature française le syntagme le plus souvent répété est *pouvoir être* mais l'on en rencontre en moyenne que 4,6 pour 10 000 mots, soit une fréquence 2,8 fois moins forte que *vouloir dire* chez les présidents. Cet écart n'est pas accidentel : on le retrouve à peu près, pour chaque ligne entre la partie gauche et la partie droite du tableau.

Quelques remarques peuvent éclairer les principales conclusions à tirer de ce tableau.

Premièrement, même avec les opérations de regroupement permises par la notion de syntagme, le phénomène reste diffus, donc difficile à observer. Il faut utiliser un "grain" plus fin. L'échelle devient 10 000 mots, et ce grossissement ne permet d'observer que le haut de la distribution, car l'essentiel des syntagmes n'apparaissent qu'un petit nombre de fois. Même dans des corpus très étendu, la majorité d'entre eux est trop rare pour faire l'objet d'un test statistique.

Deuxièmement, on n'est pas surpris par la présence de certains syntagmes comme :

- "Mesdames et messieurs" qui est la manière normale de s'adresser à une assemblée – quand il s'agit d'une allocution radio-télévisée, elle commençait autrefois par "Françaises, Français" et maintenant par "Mes chers compatriotes". *Monsieur le président* (15e rang) est le complément de cette formule d'ouverture. En effet, dans la plupart des assistances auxquelles s'adresse le président, se trouvent un ou plusieurs autres présidents (de la commission ou du parlement européen, de l'assemblée nationale, du sénat, du conseil constitutionnel, des juridictions, des régions, départements, universités, conseils, associations…). Cette omni-

présence marque jusqu'à la caricature la prégnance du modèle présidentiel (et masculin) dans notre société. Naturellement, le président *est heureux* de rencontrer ainsi d'autres présidents.

- "président de la République" et "chef de l'Etat" indiquent que – en dehors des formules de politesse - le rôle du président et ses fonctions sont le thème le plus souvent abordé. Au fond, les chefs politiques sont placés dans une position telle qu'il n'est pas nécessaire de dire *je* pour parler de soi.

Troisièmement, dans les études portant sur les collocations en français, les verbes sont très peu présents, du fait principalement de leurs nombreuses flexions qui font tomber la plupart d'entre eux dans les profondeurs des listes hiérarchiques. Le passage aux syntagmes répétés leur rend… la première place. En effet, la combinaison d'un auxiliaire modal avec un verbe à l'infinitif est un phénomène très courant, au moins aussi fréquent que la combinaison d'un auxiliaire *être* ou *avoir* avec un participe passé (Labbé & Labbé 2010).

En théorie les auxiliaires modaux sont en nombre illimité. En pratique, la liste des modalités usuelles comporte moins d'une vingtaine de verbes. Les principaux auxiliaires servent à exprimer la nécessité (*falloir*), l'obligation (*devoir*), la probabilité ou le souhait – *devoir*, *souhaiter* -, la possibilité (*pouvoir*), la connaissance (*savoir*), la volonté (*vouloir*), le futur immédiat (*aller*), le passé proche (*venir*). Ces auxiliaires modaux sont associés de manière privilégiée à certains verbes d'action, d'état, de possession, de la pensée, de la communication ou de l'action. Grâce à eux on peut donner plus de poids relatif au possible, au probable, à la volonté, à l'obligation, à la connaissance. Les hommes politiques privilégient la volonté, le possible et la nécessité.

Quatrièmement, une lecture superficielle du tableau pourrait laisser croire que les hommes politiques utilisent à peu près les même syntagmes verbaux que le reste de la population (*vouloir dir*e, *devoir* et *pouvoir être*, *pouvoir faire*…). Mais, ici, le rang et la fréquence deviennent discriminants. Ils sont également discriminants entre les hommes politiques – chacun ayant ses préférences relatives. Mais surtout, ils distinguent le discours politique du reste de la bibliothèque. Les fréquences d'emploi des principaux syntagmes verbaux sont en moyenne 2,5 fois plus élevées chez les hommes politiques que dans la littérature. Cette différence se retrouve dans la comparaison avec tous les autres corpus, même avec le français oral, pourtant gros consommateur de modalités. Rappelons la conclusion essentielle de notre communication devant les journées de l'ERLA l'année dernière :

La forte propension à utiliser des modalités verbales est une caractéristique propre au discours politique français.

Cette caractéristique est propre à la France. Elle se retrouve beaucoup moins dans les corpus Québécois et surtout Canadiens. Elle donne au discours politique français un aspect tendu qui est peut-être l'une de ses principales singularités. On peut se demander si, au niveau du discours, cette tension ne compense l'effet produit par d'autres caractéristiques comme la dépersonnalisation (déjà évoquée à propos des pronoms personnels) ou le poids prépondérant du groupe nominal que l'on va maintenant aborder.

V. Un discours peu personnalisé et privilégiant le groupe nominal

A propos du vocabulaire, le rôle des catégories grammaticales a été déjà évoqué comme variable explicative des mouvements observés au niveau des fréquences des vocables. Les caractéristiques propres au discours politique – par rapport au reste de la bibliothèque - sont résumées dans le tableau 6.

Tableau 6. Densités des catégories grammaticales dans le discours politique comparé au reste de la bibliothèque (‰)*

Catégories	A-B (Bibliothèque autre que discours politique ‰)	B (Discours politique ‰)	(B-A) / A %
Verbes	163,8	147,1	-10.2
Formes fléchies	*84,6*	*88,0*	*+4.0*
Participes passés	*45,4*	*21,7*	*-52.1*
Participes présents	*4,3*	*2,7*	*-37.5*
Infinitifs	*29,5*	*34,7*	*+17.8*
Noms propres	22,8	23,0	+0.8
Noms communs	176,9	189,3	+7.0
Adjectifs	54,2	60,8	+12.3
Adj. participe passé	*9,2*	*6,2*	*-32.4*
Pronoms	139,0	109,8	-21.0
Pronoms personnels	*83,3*	*57,2*	*-31.4*
Pronoms démonstratifs	*13,3*	*15,6*	*+17.4*
Pronoms possessifs	*0,3*	*0,3*	*-16.5*
Pronoms indéfinis	*4,8*	*3,5*	*-28.0*
Pronoms relatifs	*24,2*	*24,7*	*+2.1*
Déterminants	167,0	190,4	+14.0
Articles	*109,5*	*135,2*	*+23.5*
Nombres	*21,5*	*21,9*	*+1.6*
Possessifs	*19,9*	*15,7*	*-21.0*
Démonstratifs	*7,3*	*8,1*	*+10.5*
Indéfinis	*8,9*	*9,6*	*+7.9*
Adverbes	75,6	62,3	-17.6
Prépositions	138,5	160,8	+16.2
Conjonctions	56,1	54,6	-2.8
Coordination	*31,1*	*31,7*	*+1.7*
Subordination	*25,0*	*22,9*	*-8.3*

* les totaux sont légèrement inférieurs à 1000 du fait que les mots étrangers ne sont pas inclus dans le calcul.

Le tableau se lit de la manière suivante. Dans les autres sections de la bibliothèque, on rencontre en moyenne 163,8 verbes tous les 1000 mots. Dans le discours politique, cette moyenne est de 147,1 ‰, soit 10,2% de moins que dans le reste de la bibliothèque. Les quatre lignes suivantes détaillent ces mouvements à l'intérieur de la catégorie du verbe. Dans le discours politique, la réticence la plus forte porte sur le passé, ce qui indique que ce discours est essentiellement tourné vers le présent et le futur. Seul l'infinitif fait l'objet d'une nette préférence mais cela s'explique par l'utilisation fréquente des modalités (le verbe principal étant à l'infinitif).

La faible propension à utiliser les pronoms personnels est confirmée mais surtout elle peut être reliée à une certaine réticence dans l'emploi des verbes. L'adverbe et les conjonctions de subordination suivent les mêmes mouvements. Ces quatre catégories évoluent toujours dans le même sens.

A l'inverse, le discours politique privilégie les noms, les adjectifs, les déterminants – articles et démonstratifs – ainsi que les prépositions, dont les mouvements sont également associés.

Pour résumer ces mouvements, on peut donc rassembler ces catégories en deux groupes : nominal et verbal (tableau 7). Le groupe nominal comporte les substantifs, les adjectifs, les déterminants et les prépositions. Le groupe verbal comprend les verbes, les pronoms, les adverbes et les conjonctions de subordination. Certes, le partage n'est pas absolu : on trouve des adverbes dans le groupe nominal (notamment devant l'adjectif) ; il y a des prépositions dans le groupe verbal, etc. Mais dans tous les corpus en langue française, on observe ces corrélations entre les éléments constitutifs des deux groupes.

Tableau 7. Poids des groupes verbaux (GV) et nominaux (GN) dans le discours politique et dans d'autres sections de la bibliothèque

	GV‰	Indice	GN‰	Indice
Premiers ministres Canada	273,8	100	725,6	100
Premiers ministres Québec	347,3	127	651,2	90
Présidents français	387,1	141	682,6	94
Littérature	409,2	149	586,4	81
Français oral	442,6	162	544,3	75

Le tableau 7 indique que le groupe verbal représente 273,8 pour mille mots dans les discours des Premiers ministres canadiens contre 442,6‰ dans le français oral, soit 62% de plus. Personne n'avait soupçonné qu'il puisse exister des différences aussi importantes dans une même langue. Accessoirement, le tableau confirme le choix du discours politique envers le groupe nominal. Il montre également que ce choix est encore amplifié au Canada et, dans une moindre mesure au Québec. Le discours des chefs canadiens et québécois est relativement impersonnel. Ils insistent sur le *nous* et emploient nettement moins la première personne du singulier (*je, me, moi*) que les politiciens français. Deux explications sont possibles. Cette différence pourrait tenir à la collégialité, au moins de façade, imposée par le système parlementaire, opposée à la personnalisation inhérente au système présidentiel. Certes, dans le système de Westminster, le Premier ministre concentre beaucoup de pouvoirs entre ses mains, mais c'est en tant que chef d'une majorité parlementaire qu'il se doit d'incarner au moins en paroles. De plus, la tradition parlementaire oblige à présenter les débats sous une apparence aussi dépassionnée que possible. Mais la différence importante entre les chefs du Canada et du Québec suggère également une autre explication. A Ottawa, même si les deux langues sont officielles, le cabinet travaille en anglais et ses textes sont ensuite traduits en français. Dès lors, il faudrait étudier le poids des différentes catégories grammaticales en anglais…

Comment interpréter la nette préférence des hommes politiques pour le groupe nominal et leur réticence devant le verbe ? Trois explications sont possibles.

Premièrement, dès 1950, le statisticien P. Guiraud a signalé que le nombre des substantifs et celui des verbes varient en sens inverse et que le substantif domine dans la prose abstraite. D. Mayaffre ajoute qu'une prépondérance relative du substantif marquerait un

discours orienté vers les notions, les concepts, les idées et qu'une augmentation de la densité des verbes déplacerait le centre de gravité du discours vers «les moyens de la politique» (Mayaffre, 2004). Si l'on applique cette hypothèse à l'ensemble des hommes politiques, on devrait en conclure qu'ils préfèrent donner plus de place aux idées qu'à l'action…

Deuxièmement, pour la stylistique traditionnelle, la construction nominale «présente le fait sans date, sans mode, peut-être sans aspect, sans le rattacher nécessairement à un sujet (donc à une cause), à un objet (donc à un but)» (Cressot, 1963). La linguistique moderne nuance cette proposition de la manière suivante. Le verbe (ou ses équivalents) a une double fonction : la «fonction cohésive» qui organise «en une structure complète les éléments de l'énoncé» et la fonction assertive qui «dote l'énoncé d'un prédicat de réalité», car l'élément verbal implique une référence à un ordre qui n'est plus simplement celui du discours, mais celui de la réalité (Benveniste, 1950). Si l'on adopte ces hypothèses, la préférence pour le groupe nominal permettrait aux hommes politiques d'effacer de leurs discours (au moins partiellement) les questions pour lesquelles ils n'ont pas de réponse ou qui semblent hors de leur portée. On observe un mouvement du même genre dans les discours électoraux (Labbé & Monière 2010).

Troisièmement, chez le même auteur, le passage de l'oral à l'écrit se traduit par une diminution du poids du groupe verbal et par une augmentation parallèle du groupe nominal (Labbé 2002, 2003 ; Labbé & Labbé 2007). Autrement dit, l'expression spontanée privilégie le verbe, les pronoms et les adverbes. Le passage à l'écrit amène à remplacer un certain nombre de ces verbes par des substantifs, certains adverbes par des adjectifs, à réduire l'emploi du pronom, etc. L'effort d'élaboration s'accompagne d'un mouvement d'abstraction au-delà de la perception immédiate et du monde extérieur. Le discours politique est donc soigneusement élaboré et, même quand il prend les apparences de l'improvisation, il est éloigné de la spontanéité.

Conclusions

Ces résultats sont obtenus sur des corpus qui – bien que couvrant un très grand nombre de situations et de locuteurs, et bien que comportant au total plus de 23 millions de mots - ne prétendent pas représenter le français contemporain avec la précision que pourraient donner de véritables échantillons établis par quota. Mêmes imparfaites, ces expériences permettent néanmoins de conclure que le discours politique constitue bien un genre singulier dont on est parvenu à isoler les principales caractéristiques.

Le vocabulaire politique n'est pas constitué de mots utilisés uniquement par les hommes politiques. Les principaux vocables sont communs à tous les locuteurs du français, mais, dans le discours politique, leurs fréquences d'emploi et certaines combinaisons particulières dessinent plusieurs univers dont le plus important tourne autour des institutions, du rôle du chef et de son équipe, de la mise en valeur de sa pensée et de ses décisions. Deux autres univers dominent ce vocabulaire : le pays et ses habitants – au nom desquels les hommes politiques prétendent parler grâce à la première personne du pluriel -, les problèmes économiques et sociaux.

Les autres caractéristiques sont peut-être un peu plus inattendues. Malgré la forte présence de l'énonciateur dans ses propos, le discours politique est assez dépersonnalisé avec une tonalité nettement pédagogique. Il est orienté vers le présent et le futur. Il privilégie les généralités et les abstractions plutôt que l'action.

Dans le détail, il faut distinguer entre le discours politique dans la tradition parlementaire anglo-saxonne et dans la tradition française. Le premier privilégie le *nous* et relègue le *je* en arrière-plan. Il accentue la dépersonnalisation des propos et l'orientation vers le groupe nominal. La tradition française est marquée par une affirmation plus nette du leader et, depuis les années 1970, par la domination du *je* sur le *nous*. Les propos sont plus tendus, notamment à cause d'une sur-utilisation des modalités verbales. Mais, même dans le discours politique français contemporains, les propos restent relativement dépersonnalisés et abstraits par rapport aux autres collections de textes dont on dispose dans cette langue.

A posteriori, ces conclusions pourront sembler banales mais l'étaient-elles a priori ? Au fond, c'est le principal intérêt de ce genre d'études. Les expériences sont reproductibles, les résultats vérifiables. Les conclusions ne sont pas forcément nouvelles, mais elles changent de statut.

Enfin, la constitution de vastes corpus – comme celui du discours politique - est une nécessité pour le développement de l'analyse de la langue et du discours. Ces corpus ne sont utilisables que s'ils répondent à certains critères formels. Il faut notamment que les graphies aient été soigneusement corrigées et standardisées, et, pour le français, que tous les mots aient été lemmatisés. A ces deux conditions, la statistique appliquée au langage peut offrir des outils précieux non seulement pour les historiens et les politistes mais aussi pour les linguistes.

Remerciements

Les corpus du discours politique ont été constitués avec D. Monière, spécialement ceux du Canada-Québec, des premiers ministres français et de la campagne présidentielle de 2007. Depuis plus de 25 ans, Denis Monière a également participé à la mise au point des programmes et des analyses.

D. Mayaffre a fourni les textes de Pompidou, J. Savoy les textes politiques suisses, E. Najjar les discours tunisiens. D'autres personnes, trop nombreuses pour être toutes citées, ont aidé à la constitution des autres corpus constituant la bibliothèque du français contemporain.

Enfin, depuis 10 ans, nous avons eu de fructueux échanges avec l'ERLA.

Références

(Tous nos travaux cités sont consultables en ligne, notamment sur le site *Archives ouvertes* du CNRS).

Benveniste Emile (1956). "La nature des pronoms". Reproduit dans Benveniste 1966, p.251-265.

Benveniste Emile (1966 & 1970). *Problèmes de linguistique générale*. Paris, Gallimard (rééd. 1980).

Blumenthal Peter & Hausmann Franz J. Eds (2006). "Collocations, corpus, dictionnaires". *Langue française*, 150, juin 2006.

Cressot Marcel (1963). *Le style et ses techniques*. Paris : PUF (1ère édition : 1947).

Gougenheim Georges, en collaboration avec Michea René, Rivenc Paul, Sauvageot Aurélien (1956). *L'élaboration du français élémentaire : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris : Didier. Réédition augmentée en 1964 sous le titre : *L'élaboration du français fondamental. Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris : Didier.

Gougenheim Georges (1958). *Dictionnaire fondamental de la langue française*. Paris : Didier. Nouvelle édition revue et augmentée, Didier, Paris, 1977.

Guiraud Pierre (1950). *Les caractères statistiques du vocabulaire*. Paris : PUF.

Guiraud Pierre (1960). *Problèmes et méthodes de la statistique linguistique*. Paris : PUF.

Hubert Pierre & Labbé Dominique (1995). "La structure du vocabulaire du général de Gaulle". Communication aux 3e journées internationales d'analyse des données textuelles. In Bolasco Sergio et al. *IIIe Giornate internazionali di Analisi Statistica dei Dati Testuali*. Rome : Centro d'Informazione e stampa Universitaria, 1995, tome II, p. 165-176.

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

Labbé Cyril & Labbé Dominique (2007). "Baudelaire, Rimbaud et Verlaine". Communication aux VIIIe Journées de l'ERLA. Brest : 16-17 novembre 2007. Publié dans Banks David (Ed). *Aspects diachroniques du texte poétique*. Paris : l'Harmattan, 2011, p. 17-45.

Labbé Cyril & Labbé Dominique (2009). "Existe-t-il un genre épistolaire ? Hugo, Flaubert et Maupassant". Communication aux Xe Journées de l'ERLA. Brest : 20-21 novembre 2009.

Labbé Cyril & Labbé Dominique (2010). "La modalité verbale en français contemporain. Les hommes politiques et les autres". Communication aux XIe Journées de l'ERLA. Brest : 19 novembre 2010.

Labbé Dominique (1990a). *Le vocabulaire de F. Mitterrand*. Paris : Presse de Sciences-Po.

Labbé Dominique (1990b). *Normes de saisie et de dépouillement des textes politiques*. Grenoble : Cahier du CERAT.

Labbé Dominique (1997)."Le «nous» du général de Gaulle". *Quaderni di studi linguistici*. 4/5, 1998, p 331-354.

Labbé Dominique (2002). «Le général de Gaulle en campagne ». Communication aux IIIe Journées de l'ERLA. Reproduit dans Banks David. (ed.) *Aspects linguistiques du texte de propagande*. Paris : L'Harmattan, 2005, 213-233.

Labbé Dominique (2003). "Coordination et subordination en français oral". IVe journées de l'ERLA Coordination/subordination dans le texte de spécialité. Brest : 14-15 novembre 2003. Reproduit dans : Banks David (Ed.). *La coordination et la subordination dans le texte de spécialité*. Paris : L'Harmattan, 2007, p. 161-182.

Labbé Dominique et Monière Denis (2003). *Le vocabulaire gouvernemental. Canada, Québec, France (1945-2000).* Paris : Champion.

Labbé Dominique et Monière Denis (2010). "Quelle est la spécificité des discours électoraux? Le cas de Stephen Harper". Canadian Journal of Political Science / Revue canadienne de science politique, 43:1, (March/ mars 2010), p. 69–86.Labbé Cyril & Labbé Dominique (1994). Que mesure la spécificité du vocabulaire ? Grenoble : CERAT, décembre 1994 et juin 1997. Reproduit dans *Lexicometrica*. 3, 2001.

Labbé Dominique et Monière Denis (2008). *Les mots qui nous gouvernent*. Montréal : Monière-Wollank Editeurs, 2008.

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

Monière Denis, Labbé Cyril et Labbé Dominique (2005). "Les particularités d'un discours politique : les gouvernements minoritaires de Pierre Trudeau et de Paul Martin au Canada". *Corpus*, 4, p.79-104.

Pibarot André, Picard Jacques et Labbé Dominique (1998), "Les syntagmes répétés dans l'analyse des commentaires libres", in Mellet Sylvie (ed), 4e Journées d'analyse des données textuelles, Nice, 1998, p 507-516.

Savoy Jacques (2010). "Discours électoral et discours présidentiel". In Bolasco Sergio et al. (Eds). *Proceedings of 10th International Conference Statistical Analysis of Textual Data*. Rome : Edizioni Universitarie di Lettere Economia Diritto, Vol 2, p. 827-838.

Annexes

1. Le discours politique dans la bibliothèque du français moderne
(au 4 novembre 2011)

	Dates	Nombre de discours	Longeur (Mots)	Vocables
J. Chirac*	1995-2007	81	224 326	6 392
C. De Gaulle*	1958-1969	79	204 242	6 543
F. Mitterrand*	1981-1988	68	305 217	7 751
G. Pompidou**	1969-1974	122	216 809	7 651
N. Sarkozy président***	2007-2011	725	2 297 881	19 026
Discours premiers ministres	1945-2010	52	288 526	7 952
Présidentielle 2007	2007	132	809 384	13 653
Total France		1 259	4 339 336	
Discours Trône (Canada)	1945-2010	53	184 012	5 948
Discours Trône (Québec)	1867-2009	128	309 237	8 262
Premiers ministres Canada	1995-2010	573	837 336	14 515
Premiers Ministres Québec	1906-2010	1 032	2 645 589	22 356
Total Canada/Québec		1 786	3 927 925	
Autres pays francophones	1987-2008	85	581 683	9 866
Total discours politique		3 130	8 818 944	38 096

* Interventions radiotélévisées et conférences de presse
** Corpus en cours de constitution (textes remis par D. Mayaffre)
*** Toutes les interventions mises en ligne sur le site de la présidence.

Corpus	Dates	N auteurs	N documents	Longueur (mots)	Vocabulaire
Discours politique	1867-2011	nd	3 130	8 818 944	38 096
Français oral	XXe	nd*	357	2 730 134	18 431
Littérature :	XVIIe-XXe	56 auteurs	362**	8 419 023	52 668
Presse économique	1994-2000	7 journaux	1 166	1 557 403	33 035
Presse syndicale	1998-2000	5 journaux	335	216 415	8 452
Romans policiers	XXe	10 auteurs	24	548 680	17 974
Science économique	XXe	76 auteurs	112	645 341	13 965
Science hydrologique	XXe	nd	23	129 173	7198
Total			5 509	23 065 121	91 985

* 354 entretiens et 3 enquêtes d'opinion (réponses aux questions ouvertes)
** romans, pièces de théâtres, recueils de poésies, de nouvelles et de correspondance

2. Vocabulaire caractéristique du corpus discours politique comparé au reste de la bibliothèque du français moderne (extraits)

2 .1 Vocables significativement suremployés au seuil de 1 pour cent
(Classement par catégories grammaticales et spécificité décroissante)

Noms propres : France, Québec, Canada, Europe, Français, Canadien, Québécois, Etats-Unis, Afrique, Amérique, Montréal, Ottawa, Allemagne, Chine, Suisse, Union Européenne, Nations Unies…
Verbes : pouvoir, vouloir, devoir, falloir, permettre, travailler, comprendre, créer, vivre, assurer, souhaiter, continuer, proposer, décider, engager, aider, accepter, poser, compter, respecter…
Substantifs : gouvernement, pays, monsieur, monde, état, travail, question, politique, ministre, droit, entreprise, président, loi, emploi, développement, problème, société, besoin, économie, jeune…
Adjectifs : grand, nouveau, économique, social, national, public, français, dernier, important, politique, fédéral, international, canadien, meilleur, européen, québécois, possible, financier…
Pronoms : qui, nous, cela, celui, dont, lequel, chacun, leur, quel, certain, nôtre, ceci, eux-mêmes, soi, tel, aucun, nous-même, nul, quiconque, vous autres.
Adverbes : aussi, aujourd'hui, beaucoup, ailleurs, également, pourquoi, d'abord, seulement, ensemble, notamment, simplement, davantage, autant, lors, partout, particulièrement, demain, d'accord …
Déterminants : le, ce, notre, tout, leur, mille, premier, autre, même, certain, soixante, trente, aucun, tel, cinquante, plusieurs, quarante, deuxième, onze, troisième, vingt-et-unième, cinquième, cet…

2.2 Vocables significativement sousemployés au seuil de 1 pour cent
(Classement par catégories grammaticales et spécificité décroissante)

Noms propres : Julien, Artagnan, HLM, EDF, MATIF, Lucrèce, Murad, Hamelin, Grimoald, Duboeuf, Saint-Lazare, Lucresse, Bouygues, OPAC, Malthus, Photin, Hurel, Comaing, Candie…
Verbes : paraître, descendre, marcher, aimer, monter, mourir, rire, dire, douter, régaler, réapparaître, charmer, crever, redouter, juger, demeurer, ébahir, remontrer, poignarder, manger, sembler…
Substantifs : feu, âme, bruit, bras, amant, coeur, sang, heure, homme, coup, cheveu, voix, fille, oeil, eau, dieu, franc, soleil, ombre, bouche, fils, air, roi, mer, client, ciel, visage, rue, amour…
Adjectifs : beau, doux, vieux, petit, noir, blanc, incisif, frêle, neuf, malpropre, fiancé, complaisant, gros, étincelant, pointu, nocturne, enchanté, magnanime, ébloui, aisé, insensible, ravissant…
Pronoms : ça, ils, se, tu, vous, moi, je, on, le, il, lui, quoi, rien, toi, en, tout, quelqu'un, mien, tien, toi-même, lui-même, nous autre, eux-autres, sien, celui-là, vous-même, moi-même, vous autre…
Adverbes : point, enfin, fort, oui, là, puis, si, bon, peut-être, peu, non, trop, mal, ne, bien, loin, doucement, lentement, assez, au-dessus, toujours, soudain, presque, aussitôt, tout, brusquement
Déterminants : son, ton, mon, quelque, quel, un, votre, quatre, trois, vingt, quinze, seize, zéro, huit, six, second, dix, dix-huitième, treize, tout, quatorze, nul…