# Le vocabulaire caractéristique du Premier ministre du Québec J. Charest comparé à ses prédécesseurs.

Denis Monière, Dominique Labbé

# Le vocabulaire caractéristique du Premier ministre du Québec J. Charest comparé à ses prédécesseurs

Denis Monière[1], Dominique Labbé[2]

[1] Université de Montréal (denis.moniere@umontreal.ca)

[2] Institut d'Etudes Politiques de Grenoble (dominique.labbe@iep-grenoble.fr)

**Abstract**

How to measure the singularities of the vocabulary of an author? We answer with the help of an example: the speeches of Jean Charest (Premier of Quebec since 2003). In order to minimize the influence of the context of enunciation, we use a reference corpus containing the speeches by his predecessors (the prime ministers of Quebec since the early twentieth century). The calculation of the specific vocabulary is presented, several improvements are proposed. In addition to the characteristics of speeches by J. Charest, this experience shows that the more a word is used, the more likely it is to be specific.

**Résumé**

Comment mesurer les singularités du vocabulaire d'un locuteur ? On répond à l'aide d'un exemple : les discours de Jean Charest (Premier ministre du Québec depuis 2003). Pour minimiser l'influence du contexte de l'énonciation, on utilise un corpus de référence contenant les discours émis par ses prédécesseurs (les Premiers ministres du Québec depuis le début du XXe siècle). Le calcul des spécificités du vocabulaire est rappelé, plusieurs améliorations sont proposées. Outre les caractéristiques des discours de J. Charest, l'expérience montre que plus un vocable est employé, plus il a de chances d'être spécifique.

**Mots-clés :** Discours politique ; corpus ; statistique lexicale ; spécificités du vocabulaire ; catégories grammaticales ; Québec ; Charest.

## 1. Introduction

Une vaste collection de textes a été rassemblée dans une "grande bibliothèque du français contemporain" (voir annexe 1). Cette bibliothèque permet de renouveler l'étude de la langue, de l'histoire littéraire comme celle du discours politique. Grâce à elle, on peut enfin répondre à des questions simples comme : quel est le vocabulaire (ou le style) caractéristique d'un locuteur. Pour illustrer cette idée, on utilisera les discours prononcés par l'actuel Premier ministre du Québec (J. Charest). Les 174 interventions, qu'il a prononcées jusqu'au 31 décembre 2010, figurent parmi les 3 130 discours – d'une centaine de locuteurs différents – de la section "discours politiques" de la bibliothèque électronique. Cette section comprend plus de 8,8 million de mots, dont près de la moitié proviennent des discours des Premiers ministres du Québec et du Canada (Labbé & Monière 2003 ; Labbé & Monière 2008).

Après avoir rappelé la manière dont ces corpus sont constitués, on examinera la question suivante : comment déterminer les particularités du vocabulaire du Premier ministre actuel par rapport à celui de ses 12 prédécesseurs ?

## 2. Les corpus étiquetés

Avant son entrée dans la bibliothèque, chaque texte subit une série de traitements. Outre l'ajout des références (auteur, titre, lieu, date… comme dans un catalogue), l'orthographe est corrigée et les graphies sont standardisées. Puis 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. Voici un exemple, tiré du premier discours de Jean Charest (29 avril 2003)

présentant son nouveau ministère :

"Notre cabinet est réduit. Il compte 18 ministres en titre."

Ces deux phrases comportent plusieurs ambiguïtés :
- *est* : verbe "être" ou substantif masculin (point cardinal) ?
- *réduit*, verbe "réduire" (indicatif ou participe passé), substantif masculin ("un réduit") ?
- *compte* : verbe "compter" ou substantif masculin ("un compte") ?
- *ministre* : substantif masculin ou féminin ?
- *en* : pronom ou préposition ?
- *titre* : verbe "titrer" ou substantif masculin singulier ("un titre") ?

Soit 6 difficultés pour 10 mots. Cette proportion ne doit pas surprendre : dans tout texte français, en moyenne, plus du tiers des mots sont homographes (une seule graphie et deux ou plusieurs entrées de dictionnaire). Ces homographies sont résolues par un analyseur syntaxique qui attache à chaque mot une étiquette indiquant l'entrée de ce mot dans les dictionnaires de langue. On désigne cette opération sous le nom de "lemmatisation" mais "étiquetage" serait préférable. Le tableau 1 ci-dessous donne un exemple d'étiquetage.

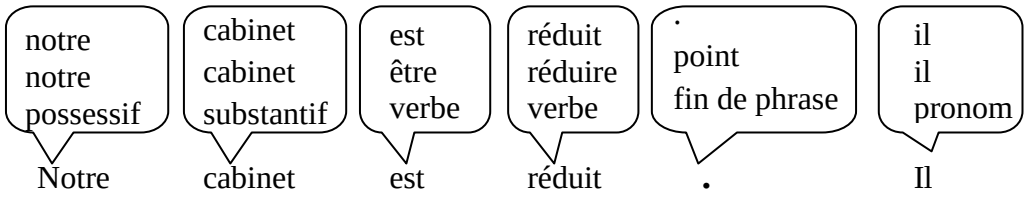



**Tableau 1**. *Etiquetage d'une phrase de J. Charest*

Dans l'étiquette, "notre" est la graphie standard, la seconde ligne est l'entrée de dictionnaire et la troisième, la catégorie grammaticale. Les conjugaisons d'un verbe sont groupées sous son infinitif ou les pluriels du substantif sous le singulier ou encore le féminin et le pluriel de l'adjectif sous le masculin singulier (Labbé, 1990). Ces traitements ne modifient pas le texte original. Les étiquettes s'y ajoutent et sont autant de portes d'entrée dans le texte, comparables aux entrées d'un dictionnaire.

Ces traitements offrent de nombreux intérêts, spécialement pour l'étude de la langue et du vocabulaire. Ils permettent de déterminer les particularités du discours de J. Charest, grâce au calcul dit des "spécificités du vocabulaire".

## 3. Les spécificités du vocabulaire

Pour ce calcul, la loi normale semble la solution appropriée : on considère les discours de J. Charest comme autant d'échantillons extraits aléatoirement, et avec remise, dans une urne de Bernouilli constituée par le corpus de référence (pour une application récente : Savoy, 2010). Le calcul est simple et ses résultats significatifs, mais ce schéma peut se voir opposer plusieurs objections. Par exemple, pour que le prélèvement n'affecte pas le contenu de l'urne, il faudrait que les échantillons soient petits par rapport à la dimension de l'urne. Les corpus disponibles rendent ce pré-requis irréaliste. Par exemple, le Corpus "premiers ministres du Québec" compte 2 645 591 mots et l'ensemble des discours de J. Charest : 300 068. Il est impossible de considérer que le "prélèvement" du second sera sans influence sur la composition du premier (l'urne). De plus, tout texte en langue naturelle comporte une proportion considérable de mots de faible fréquence : aussi grand que soit le corpus de référence, il contiendra toujours une majorité d'*hapax* et de vocables apparaissant très rarement. Le tirage d'un de ces mots "rares" aura une influence évidente sur le contenu de

l'urne et sur les épreuves suivantes.

Il est donc nécessaire d'utiliser la loi hypergéométrique — ou “tirage sans remise” (le tirage d'un vocable modifie son espérance mathématique de figurer dans les tirages suivants) — et d'inclure explicitement dans l'urne le corpus sous revue. Le calcul présenté ci-dessous est inspiré de celui proposé par P. Lafon pour les “spécificités du vocabulaire” (Lafon, 1984) reformulé par Labbé & Labbé (1994) et appliqué pour la première fois par Hubert & Labbé (1995) sur le vocabulaire du général de Gaulle.

Soit :
— le corpus de référence (*C*) composé de $N_c$ occurrences (longueur en “mots”) ;
— le sous-corpus étudié (*B* avec $B \subset C$) composé de $N_b$ occurrences ;
— un vocable *i* avec $F_{ic}$ occurrences dans *C* et $F_{ib}$ dans *B*.

Si les mêmes lois de composition sont à l'œuvre dans la population totale *C* et dans la sous-population *B*, alors l’effectif théorique d'un vocable *i* dans *B* ($E_{ib}$) sera le nombre de ses occurrences dans *C* pondéré par le rapport entre la taille de *B* et celle de *C* :

$$E_{ib(u)} = F_{ic} * U \; avec \; U = \frac{N_b}{N_c} \qquad (1)$$

Si le nombre d’occurrences constatées ($F_{ib}$) est différent de celui attendu ($E_{ib}$), peut-on dire que le vocable est significativement sur-employé ou sous-employé dans *B* par rapport à *C* ? Pour répondre à cette question, il faut considérer la probabilité de l'événement observé $F_{ib}$ par rapport à l'événement attendu ($E_{ib}$). Cette probabilité est la combinaison de deux événements :

— le nombre de possibilités différentes de choisir $N_b$ mots dans un total de $N_c$:

$$C_c^b = \frac{N_c!}{N_b!(N_c - N_b)!} = \begin{bmatrix} N_c \\ N_b \end{bmatrix}$$

— le nombre de possibilités différentes de choisir $F_{ib}$ mots dans un total de $F_{ic}$ :

$$C_{F_{ic}}^{F_{ib}} = \frac{F_{ic}!}{F_{ib}!(F_{ic} - F_{ib})!} = \begin{bmatrix} F_{ic} \\ F_{ib} \end{bmatrix}$$

La probabilité composée de ces deux événements suit une loi hypergéométrique dont les paramètres sont : $F_{ic}$, $F_{ib}$, $N_b$, $N_c$:

$$P(X = F_{ib}) = \frac{\begin{bmatrix} F_{ic} \\ F_{ib} \end{bmatrix}\begin{bmatrix} N_c - F_{ic} \\ N_b - F_{ib} \end{bmatrix}}{\begin{bmatrix} N_c \\ N_b \end{bmatrix}} \qquad (2)$$

L’indice de spécificité (S) est la somme des probabilités – calculées avec (2) – de survenue des *J* valeurs entières de *X* variant de 0 à $F_{ib}$ {*X*=0 ; *X*= $F_{ib}$} :

$$S = P(X \leq F_{ib}) = \sum_{j=0}^{j=Fib} P(X = j) \qquad (3)$$

On choisit $\alpha$ un seuil de risque d’erreur de première espèce (1% dans la suite de ce travail). Si *S* est inférieur à ce seuil,
- avec $F_{ib} < E_{ib}$, le vocable *i* est une "spécificité négative" (il serait plus exact de dire que *i* est significativement sous-employé au seuil choisi) ;
- avec $F_{ib} > E_{ib}$ le vocable *i* est une "spécificité positive" (*i* est significativement sur-employé).
Par exemple, J. Charest emploie au total 8 219 verbe *être* alors qu’on en attend 8 200 ($E_{ib(u)}$). Au seuil choisi, la différence n’est pas significative. On en déduit que J. Charest n’utilise

probablement pas ce verbe différemment de ses collègues. En revanche, il a utilisé 6 179 fois le verbe *avoir*, alors que l'espérance mathématique était de 6 748. Au seuil de 1%, J. Charest aurait significativement moins employé ce verbe que ses prédécesseurs.

Les résultats de ce calcul semblent pourtant contestables.

*3.2 Singularités du calcul appliqué à J. Charest*

Appliquée aux corpus Charest - comparé au corpus de l'ensemble des premiers ministres québécois -, la formule (2) donne les résultats suivants :
- une majorité de verbes se trouvent en spécificités négatives (sous-emplois), notamment parmi les plus fréquents : *avoir, pouvoir, dire, vouloir, devoir, falloir, croire, savoir*…
- il en est de même pour la plupart des pronoms, notamment les pronoms personnels sauf *nous* et *vous* – et des adverbes (spécialement *ne*, *pas* et *plus*) ;
- en revanche, il y a plus de substantifs dans les spécificités positives (sur-emplois) que dans les spécificités négatives. De même, les principaux articles *(le, un*…) et prépositions (*de, à, pour, par*) seraient significativement sur-employées par J. Charest.
Ces déséquilibres peuvent s'expliquer en partie par certaines caractéristiques du vocabulaire de J. Charest (tableau 2).

| Catégories | A (Premiers ministres - Charest) ‰ | B (Charest) ‰ | B-A (%) |
|---|---|---|---|
| Verbes | 147,6 | 137,7 | -6,7 |
| *Formes fléchies* | *89,7* | *84,0* | *-6,4* |
| *Participes passés* | *22,1* | *20,0* | *-9,7* |
| *Participes présents* | *2,4* | *2,4* | *+2,1* |
| *Infinitifs* | *33,5* | *31,4* | *-6,1* |
| Noms propres | 21,1 | 32,5 | +54,0 |
| Noms communs | 185,0 | 201,3 | +8,8 |
| Adjectifs | 60,8 | 56,2 | -7,7 |
| *Adj. participe passé* | *5,7* | *5,1* | *-10,0* |
| Pronoms | 114,1 | 102,0 | -10,6 |
| *Pronoms personnels* | *59,3* | *53,8* | *-9,3* |
| *Pronoms démonstratifs* | *15,1* | *14,4* | *-4,8* |
| *Pronoms possessifs* | *0,3* | *0,1* | *-51,4* |
| *Pronoms indéfinis* | *3,2* | *2,2* | *-31,7* |
| *Pronoms relatifs* | *25,4* | *21,8* | *-14,1* |
| Déterminants | 189,9 | 205,5 | +8,2 |
| *Articles* | *127,4* | *141,0* | *+10,7* |
| *Nombres* | *28,5* | *31,1* | *+8,9* |
| *Possessifs* | *14,6* | *16,8* | *+14,8* |
| *Démonstratifs* | *8,4* | *9,0* | *+6,9* |
| *Indéfinis* | *11,0* | *7,7* | *-30,2* |
| Adverbes | 66,6 | 50,4 | -24,3 |
| Prépositions | 158,7 | 171,2 | +7,8 |
| Conjonctions | 54,7 | 42,7 | -21,8 |
| *Coordination* | *30,3* | *27,0* | *-10,8* |
| *Subordination* | *24,4* | *15,7* | *-35,6* |

**Tableau 2.** *Densités des catégories grammaticales chez J. Charest comparé aux autres premiers ministres québécois (en ‰)*

Ce tableau se lit de la manière suivante : la première ligne signifie que, dans le corpus de référence (l'ensemble des premiers ministres sauf J. Charest), on rencontre 148 verbes en moyenne pour 1000 mots et que, dans les discours prononcés par J. Charest, cette proportion tombe à moins de 138 ‰, soit un "recul" de 6,7 %. Ce recul est particulièrement sensible pour les participes passés (-9,7 %). Seuls les participes présents y échappent. Pour les autres catégories, les écarts sont, pour la plupart, suffisamment importants pour ne pas provenir de simples fluctuations aléatoires. On remarque que, d'un côté, les pronoms, les adverbes et les conjonctions de subordination reproduisent les mouvements du verbe, de manière amplifiée et que, de l'autre, ces reculs sont compensés par une augmentation des substantifs, des déterminants et des prépositions.

Ce tableau explique certains résultats du calcul des spécificités du vocabulaire : puisqu'il y a moins de verbes dans les discours de J. Charest que dans le corpus de référence, il est logique que les principaux verbes apparaissent en spécificités négatives, avec la plupart des pronoms et des adverbes. A l'inverse, il est également logique que la majorité des substantifs, des déterminants et des prépositions figurent en spécificités positives.

*3.3 Pondération du calcul des spécificités par la densité des catégories grammaticales*

Pour pallier ce biais, on utilise la densité des catégories grammaticales pour pondérer la formule (2) ci-dessous. Cette méthode a été indiquée dans Labbé et Labbé (1994) et développée dans Monière, Labbé & Labbé (2005) et dans Labbé & Labbé (2005).

Soit : $N_{gb}$ et $N_{gc}$ la somme des effectifs de tous les vocables appartenant à la catégorie grammaticale *G* dans le sous-corpus *B* et respectivement dans le corpus entier *C*. Afin de donner à un vocable *i* une chance égale — en fonction de ses effectifs totaux dans le corpus entier - et indépendante de sa catégorie grammaticale *G*, d'intervenir dans le sous-corpus étudié, les formules ci-dessus (1) et (2) deviennent :

$$E_{ib(u)} = F_{ic} * \frac{N_{gb}}{N_{gc}} \qquad (4)$$

$$P(X = F_{ib}) = \frac{\begin{bmatrix} F_{ic} \\ F_{ib} \end{bmatrix} \begin{bmatrix} N_{gc} - F_{ic} \\ N_{gb} - F_{ib} \end{bmatrix}}{\begin{bmatrix} N_{gc} \\ N_{gb} \end{bmatrix}} \qquad (5)$$

La suite du raisonnement est le même. Comme le remarque (Mayaffre 2006), cela revient à considérer le corpus, non comme une seule urne, mais comme autant d'urnes qu'il y a de catégories grammaticales, et le sous-corpus comme le résultat d'autant de tirages dans ces différentes urnes.

Les formules (4) et (5) appliquées au corpus Charest, comparé à celui de tous les autres premiers ministres québécois, aboutissent à des effectifs sensiblement égaux pour les vocables significativement sur- et sous-employés, ce qui est plus satisfaisant que les résultats obtenus avec les formules (1) et (2). Par exemple, cela change le classement des verbes *être*, *avoir* ou *vouloir* cités plus haut. Si J. Charest avait utilisé autant le verbe *être* que ses prédécesseurs, ses discours devraient en contenir 7 709 (et non pas 8 200 comme avec la formule 1). Les effectifs réels sont de 8 219. Cet écart est significatif : la sur-utilisation du verbe *être* est une caractéristique de J. Charest (alors que le calcul standard amenait à conclure à un écart non significatif). A l'inverse, le verbe *avoir* est attendu 6 321 fois, ce qui ne s'écarte pas significativement – au seuil de 1 % - des 6 179 que l'on rencontre effectivement, contrairement à ce que laissait penser le calcul standard (qui concluait à la sous-utilisation). La majorité des verbes usuels se trouvent ainsi reclassés.

Le vocabulaire caractéristique de J. Charest est résumé en annexe 2. Il illustre également sa préférence pour le nom.

## 4. La préférence de J. Charest pour le nom

Le vocabulaire d'un locuteur est maintenant caractérisé par des choix de deux ordres : d'une part, le choix en faveur – ou en défaveur – de telle ou telle catégorie grammaticale ; d'autre part, au sein de chacune de ces catégories, le choix pour tel ou tel vocable.

Pour résumer ces mouvements, on rassemble ces catégories en deux groupes : nominal et verbal. Le premier comporte les substantifs, les adjectifs, les déterminants et les prépositions. Le second, 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. Cette réserve admise, le tableau 3 confirme le choix de J. Charest envers le groupe nominal. Il montre également que ce choix est, dans une moindre mesure, partagé par la moyenne de ses collègues les Premiers ministres depuis un siècle, par rapport à l'ensemble de référence (la bibliothèque).

| | Effectifs | Proportion (‰)* |
|---|---|---|
| Charest | | |
| Groupe nominal | 208 125 | 693.6 |
| Groupe verbal | 91 746 | 305.8 |
| Premiers ministres québécois | | |
| Groupe nominal | 1 722 768 | 651.2 |
| Groupe verbal | 918 745 | 347.3 |
| Français moderne | | |
| Groupe nominal | 14 052 091 | 616.0 |
| Groupe verbal | 8 659 146 | 379.7 |

**Tableau 3.** *Poids du groupe nominal dans les discours de J. Charest comparé à l'ensemble de ses prédécesseurs et à l'ensemble de la bibliothèque.*

Comment interpréter la nette préférence de J. Charest pour le groupe nominal et ses réticences devant le verbe ? Quatre 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 (2004) 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». Si l'on applique cette hypothèse au discours de J. Charest, on devrait en conclure qu'il a choisi de donner plus de place aux idées qu'à l'action. Cela paraît cadrer assez bien avec la position prudente et attentiste adoptée par le premier ministre depuis plusieurs années.

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 ajoute que le verbe (ou ses équivalents) assure la «cohésion de l'énoncé» et le dote d'un «prédicat de réalité» (Benveniste, 1950). Dans cette optique, la préférence pour le groupe nominal permettrait à J. Charest d'effacer de son discours (au moins partiellement) les questions pour lesquelles il n'a pas de réponse ou qui semblent hors de portée de son gouvernement. 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). Autrement dit, l'expression spontanée privilégie le verbe, les pronoms, les adverbes. Le passage à l'écrit amène à remplacer un certains verbes par des substantifs, des adverbes par des adjectifs, à réduire l'emploi du démonstratif, etc. On constate en effet que, depuis plusieurs années, le Premier ministre raréfie ses conférences de presse et les rencontres improvisées et se contente le plus souvent de courtes déclarations.

Quatrièmement, chez J. Charest, on note la forte densité des participes présents. Cressot (1963) a signalé les caractéristiques particulières du participe présent, forme verbale la plus proche de l'adjectif : «Cette forme a pris au XIXe siècle un développement considérable, surtout à partir de Flaubert. Les écrivains qui attribuent aux choses une vie et une volonté secrètes, ont compris l'utilité de l'adjectif verbal pour leur expression dynamique du monde, et la possibilité d'atténuer, grâce à lui, la note trop éclatante des adjectifs en –eur et en –teur ».

Enfin, comme indiqué dans Labbé & Monière (2008), il faut traiter à part les mots à majuscule initiale – patronymes, toponymes, sigles des organisations… - et les chiffres. Ce sont des sortes d'interfaces entre le discours et la réalité extérieure à celui-ci. Les mots à majuscules assurent l'ancrage des propos dans l'espace géographique, économique, social ; les dates ancrent le discours dans le temps, les chiffres offrent des représentations des phénomènes économiques, etc. Plus la densité de ces catégories est importante, plus le discours prend une apparence concrète. Or, le tableau 2 indique que J. Charest utilise beaucoup plus de mots à majuscule et de chiffres que ses prédécesseurs. Cela suffit-il à compenser la faible personnalisation, l'abstraction relative de ses propos et à lui assurer l'image d'un politicien pragmatique à laquelle il tient tant ?

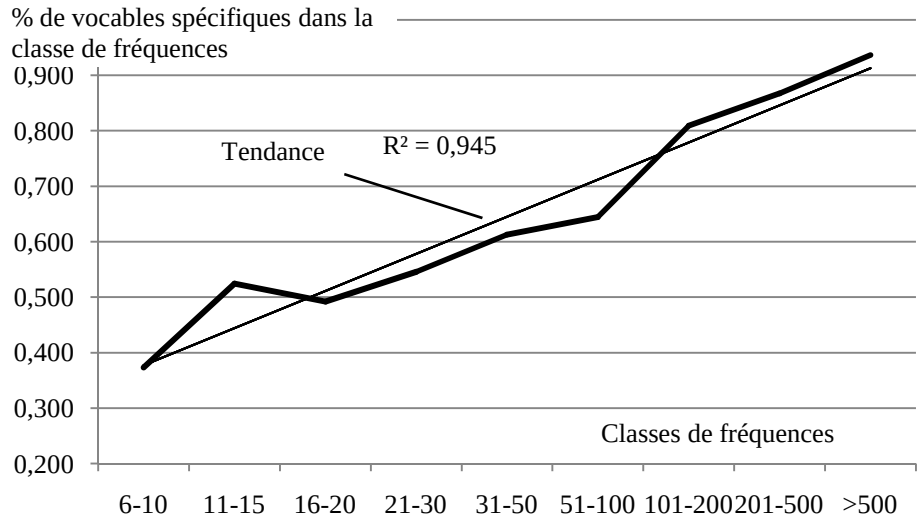
## 5. Les limites du calcul

Dans le vocabulaire de J. Charest, il y a 2 763 vocables pour lesquels le calcul des spécificités a une pleine signification (le nombre d'occurrences dans le corpus et dans le sous-corpus donne à chacun d'eux sensiblement autant de chances d'être S+ que S-, avec $\alpha = 1\ \%$). Parmi eux, 1 486, soit 53,8 % sont spécifiques à ce Premier ministre (tableau 4). Si le phénomène était seulement régi par le hasard, cette proportion ne devrait être que de 1 %. L'hypothèse nulle sur laquelle se fonde le raisonnement – les mots de J. Charest sont tirés au hasard dans le vocabulaire politique - peut être rejetée sans aucun risque d'erreur !

Les proportions, en dernière colonne du tableau 4, interrogent. Quand un vocable est utilisé plus de 500 fois par J. Charest, il a 94% de chances de lui être spécifique (en plus ou en moins). En revanche, pour un vocable utilisé entre six et dix fois, cette probabilité n'est plus que de 37,3%. Les chiffres de la dernière colonne augmentent régulièrement (graphique 1).

| Nombre d'occurrences du vocable | Nombre de vocables spécifiques dans la classe | Nombre total de vocables dans la classe | Proportion des vocables spécifiques dans la classe |
|---|---|---|---|
| 6-10 | 321 | 861 | 0,373 |
| 11-15 | 239 | 456 | 0,524 |
| 16-20 | 123 | 250 | 0,492 |
| 21-30 | 162 | 297 | 0,545 |
| 31-50 | 166 | 271 | 0,613 |
| 51-100 | 183 | 284 | 0,644 |
| 101-200 | 148 | 183 | 0,809 |
| 201-500 | 85 | 98 | 0,867 |
| >500 | 59 | 63 | 0,937 |
| Total | 1 486 | 2 763 | 0,538 |

**Tableau 4.** *Proportion de vocables spécifiques selon leur nombre d'occurrences*




**Graphique 1.** *Proportion des vocables spécifiques selon les classes de fréquences (trait gras) et ajustement linéaire (trait maigre)*

L'origine des ordonnées n'est pas à zéro, ce qui accentue les écarts entre les valeurs observées et les valeurs ajustées. Le coefficient de détermination ($R^2$) – du % de vocables spécifiques par la fréquence absolue – est pratiquement égal à 1. Autrement dit, plus un vocable est employé, plus il a de chance d'être spécifique. Cette liaison quasi-rigide est toujours observée dès que le sous-corpus et le corpus de référence comportent respectivement au moins 50 000 mots et 500 000 mots. Le calcul des spécificités enregistre surtout le fait que le vocable analysé est plus ou moins usuel et non pas d'abord l'effet de choix stylistiques ou thématiques, comme on le croit habituellement. Cela a été expliqué en conclusion de Labbé & Labbé (1994). Le tableau 5 illustre le phénomène avec les vocables les plus employés par J. Charest.

Ces 20 vocables donnent 137 468 mots, sur un total de 300 068, c'est-à-dire qu'ils couvrent à eux seuls 46% de la surface du texte. Parmi eux, seuls trois sont « non-spécifiques », avec $\alpha = 1$ % (au seuil de 5 %, *avoir* est S-). De plus, la plupart des indices sont infinitésimaux (dernière colonne). Rappelons qu'un exposant 10 signifie qu'il y 9 zéros avant le premier chiffre significatif. En fait, « 1.e-10 » est donné ici pour indiquer que la probabilité est inférieure ou égale à 0,0000000001. En effet, tous les chiffres plus petits excèdent la précision

du calcul effectué par l'ordinateur.

| Rang | Vocable | Effectifs | Spécificités |
|---|---|---|---|
| 1 | le (det) | 36 521 | S+ (8,1.e-3) |
| 2 | de (pré) | 27 196 | S- (6,7.e-3) |
| 3 | être (v) | 8 219 | S+ (1.e-10) |
| 4 | à (pré) | 8 070 | S- (1.e-10) |
| 5 | et (cj) | 6 632 | S+ (1.e-10) |
| 6 | avoir (v) | 6 179 | = |
| 7 | un (det) | 5 784 | = |
| 8 | nous (pro) | 4 324 | S+ (1.e-10) |
| 9 | ce (pro) | 3 536 | S+ (1.e-10) |
| 10 | qui (pro) | 3 358 | S+ (7,9.e-8) |
| 11 | pour (pré) | 3 338 | S+ (1.e-10) |
| 12 | en (pré) | 3 328 | S+ (1.e-10) |
| 13 | Québec | 2 894 | S+ (1.e-10) |
| 14 | dans (pré) | 2 869 | S+ (1.e-10) |
| 15 | que (cj) | 2 855 | S- (1.e-10) |
| 16 | ce (det) | 2 687 | = |
| 17 | notre (det) | 2 548 | S+ (1.e-10) |
| 18 | je (pro) | 2 404 | S- (1.e-10) |
| 19 | il (pro) | 2 383 | S- (1.e-10) |
| 20 | on (pro) | 2 343 | S+ (1.e-10) |
| Total | | 137 468 | |

**Tableau 5.** *Les 20 vocables les plus fréquents de J. Charest et leurs indices de spécificités*

## 6. Conclusions

Lorsqu'il communique avec les autres, tout locuteur effectue un choix fondamental en faveur du nom ou du verbe. Pour l'homme politique, ce choix traduit un certain rapport à l'action et au pouvoir. Favoriser le groupe nominal et dépersonnaliser son propos, c'est prendre une distance par rapport à l'action (qui comporte toujours un risque) ; c'est privilégier la conservation sur le changement. Il faut toutefois nuancer ce choix dichotomique : certains éléments du vocabulaire, comme les mots à majuscules ou les nombres, permettent de compenser, en tout ou partie, les effets d'une trop nette préférence pour le nom.

Le calcul des spécificités n'a donc guère de sens quand il est effectué sur les formes graphiques car il enregistre ce choix fondamental et non pas les choix en faveur de tel ou tel vocable. Naturellement, pour pouvoir pondérer le calcul par les catégories grammaticales, il faut que les textes aient été soigneusement étiquetés (lemmatisés) auparavant.

Deux précautions supplémentaires semblent nécessaires :

- dans les listes des spécificités, il est préférable de limiter les comparaisons aux vocables appartenant aux mêmes catégories grammaticales et à des classes de fréquence semblables ou proches ;

- la plupart du temps, la valeur de l'indice de spécificité n'a guère de sens et il est préférable de se limiter à enregistrer le fait que le vocable considéré est significativement sur- (ou sous-) employé. Rappelons à ce sujet que Labbé & Labbé (1994) propose un certain nombre d'améliorations utiles pour la présentation et l'interprétation des résultats du calcul.

Enfin, la constitution de vastes corpus étiquetés – comme celui du discours politique québécois et canadien - est une nécessité pour le développement de l'analyse du discours. À

ce propos, il faut souligner que ces corpus ne pourront être utilisables que s'ils répondent à certains critères formels. Il faut notamment que les graphies aient été soigneusement corrigées et standardisées, que les mêmes conventions soient respectées dans toute la bibliothèque et que la lemmatisation soit sans erreur. A ces conditions, la statistique appliquée au langage peut offrir des outils précieux non seulement pour les historiens et les politistes mais aussi pour les linguistes et les styliciens.

**Remerciements**

Cyril Labbé a écrit les programmes informatiques avec D. Labbé et il a contribué à l'élaboration de la grande bibliothèque. J. Savoy et C. Labbé ont relu une première version de ce texte et ont proposé plusieurs corrections et améliorations importantes. D. Mayaffre a fourni les textes du corpus Pompidou, J. Savoy ceux du discours politique suisse.

**Références**

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

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

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

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

Hubert P. & Labbé D. (1995). "La structure du vocabulaire du général de Gaulle". Communication aux 3e journées internationales d'analyse des données textuelles. In Bolasco S. 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 C. & Labbé D. (1994). *Que mesure la spécificité du vocabulaire ?* Grenoble: CERAT, décembre 1994 & juin 1997. Disponible en ligne sur le site de la revue *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 D. (1990). *Normes de saisie et de dépouillement des textes politiques*. Grenoble: Cahier du CERAT.

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

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

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

Labbé D. et Monière D. (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.

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

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

Mayaffre D. (2006). "Faut-il pondérer les spécificités lexicales par la composition grammaticale des textes ? Tests logométriques appliqués au discours présidentiel sous la Vème République". In Condé C. & Viprey J.-M. *Actes des 8e Journées internationales d'Analyse des données textuelles*. Besançon : Presses universitaires de Franche Comté, II, p. 677-685.

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

Savoy J. (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.

**Annexe 1**

Les Premiers ministres québécois dans le corpus du discours politiques en français (au 4 novembre 2011).

| | Dates | Nombre de discours | Longueur (mots) | Vocables différents |
|---|---|---|---|---|
| L. Gouin (libéral) | 1905-1919 | 19 | 102 462 | 5 010 |
| L.-A. Taschereau (libéral) | 1919-1936 | 27 | 77 241 | 5 056 |
| A. Godbout (libéral) | 1939-1944 | 36 | 78 449 | 5 466 |
| M. Duplessis (Union nationale) | 1936-1939<br>1944-1957 | 34 | 90 843 | 4 528 |
| J. Lesage (libéral) | 1960-1966 | 141 | 307 310 | 8 296 |
| D. Johnson (Union nationale) | 1966-1968 | 40 | 61 703 | 4 328 |
| J.-J. Bertrand (Union nationale) | 1968-1970 | 32 | 34 507 | 3 482 |
| R. Bourassa (libéral) | 1970-1976<br>1985-1993 | 126 | 372 586 | 7 439 |
| R. Lévesque (Parti québécois) | 1976-1985 | 93 | 452 653 | 10 147 |
| J. Parizeau (Parti québécois) | 1994-1995 | 42 | 140 446 | 5 976 |
| L. Bouchard (Parti québécois) | 1996-2001 | 174 | 431 939 | 10 278 |
| B. Landry (Parti québécois) | 2001-2003 | 94 | 195 382 | 7 988 |
| J. Charest (libéral) | 2003-2010 | 174 | 300 068 | 8330 |
| Total | | 1032 | 2 645 591 | 22 223 |

| | Dates | Nombre de textes | Mots | Vocables |
|---|---|---|---|---|
| France : | | | | |
| J. Chirac | 1995-2007 | 81 | 224 326 | 6 392 |
| C. de Gaulle | 1958-69 | 79 | 204 242 | 6 543 |
| F. Mitterrand | 1981-88 | 68 | 305 217 | 7 751 |
| G. Pompidou | 1969-1974 | 122 | 216 809 | 7 651 |
| Discours gouvernemental | 1945-2010 | 52 | 288 526 | 7 952 |
| Présidentielles 2007 | 2007 | 132 | 809 384 | 13 653 |
| N. Sarkozy président | 2007-2011 | 725 | 2 290 832 | 19 026 |
| Total France | | 1 259 | 4 339 336 | |
| Canada-Québec : | | | | |
| Discours du trône Canada | 1945-2010 | 53 | 184 012 | 5 948 |
| Discours du 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 | 22 355 |
| Total Canada - Québec | | 1 786 | 3 976 176 | |
| Autres pays francophones | 1987-2008 | 85 | 551 683 | 9 866 |
| Total discours politique | | 3 130 | 8 867 193 | 38 160 |

**Annexe 2** Vocabulaire spécifique de J. Charest comparé aux autres premiers ministres (Extraits - Classement par catégories grammaticales et spécificité décroissante)

**1. Vocables significativement suremployés au seuil de 1 pour cent**

Noms propres : Québec, Québécois, Montréal, France, Etats-Unis, Europe, Robert, Bourassa, Inde, Sherbrooke, Bavière, Laval, Pierre, Chine, PME, UNESCO, McGill, Claude…

Verbes : faire, aller, permettre, mettre, être, travailler, créer, passer, continuer, rappeler, annoncer, vivre, investir, présenter, recevoir, reconnaître, relever, développer, représenter, engager, changer, contribuer, réduire, remercier, souhaiter, ouvrir, appuyer, améliorer, saluer…

Substantifs : gouvernement, région, développement, année, dollar, ministre, état, projet, santé, service, travail, monde, économie, madame, entreprise, citoyen, plan, président, avenir, investissement, changement, défi, énergie, secteur, milliard, entente, responsabilité, vie...

Adjectifs : nouveau, québécois, grand, important, dernier, public, national, international, meilleur, durable, prochain, énergétique, régional, fier, fort, responsable, mondial, américain, universitaire, européen, forestier, démographique, manufacturier, stratégique, distingué…

Pronoms : nous, ce, on, vous, ça, moi, chacun

Adverbes : aussi, aujourd'hui, très, là, également, où, alors, puis, beaucoup, ici, mieux, ailleurs, ensemble, notamment, partout, davantage, justement, là-dessus, différemment, effectivement, d'emblée, juste, rapidement, par-delà, surtout, directement, correctement…

Déterminants : le, ce, notre, mille, deux, quatre, premier, votre, trois, cinq, six, chaque, deuxième, troisième, centième, quatrième, vingt-et-unième, septième, treize, quatorze,

Conjonctions et prépositions : de, pour, en, dans, sur, avec, parce que, entre, vers, afin, parmi, derrière, en-deçà, dès, depuis, envers,

**2. Vocables significativement sous-employés au seuil de 1 pour cent**

Noms propres : Ottawa, Chrétien, Trudeau, Meech, Duplessis, Taschereau, Union Nationale, Lévesque, Parizeau, Girard, Godbout, Cri, Canadien, Saint Jean, Hydro

Verbes : apporter, vérifier, donner, satisfaire, considérer, fournir, comprendre, aboutir, pouvoir, falloir, croire, savoir, dire, essayer, oublier, rester, accepter, demander, trouver, posséder, discuter, exister, examiner, sembler, décider, suffire, laisser, penser …

Substantifs : bill, assurance-chômage, comté, statistique, élément, armée, grandeur, pêcherie, pouvoir, cas, problème, province, chose, vue, formule, régime, électeur, législation, colonisation, constitution, cultivateur, parti, opinion, soir, agriculture, cause, attitude …

Adjectifs : légitime, simple, seul, intéressant, souverainiste, pire, provincial, libéral, évident, actuel, anglais, sûr, considérable, agricole, constitutionnel, définitif, matériel, vieux, prêt, conservateur, dit, certain, électoral, normal, facile, possible, industriel, fédéral, compris…

Pronoms : ils, lui, il, le, cela, je, quel, dont, se, lui-même, rien, en, eux-mêmes, que, leur, y, tout, lequel, tel, nous-même, vôtre, certain, aucun, celui, ceci, celui-là, autre, celui-ci, nôtre…

Adverbes : ne, point, bien, encore, pas, non, même, déjà, simplement, si, assez, seulement, moins, tout, peut-être, vite, plutôt, trop, peu, présentement, d'accord, mal, autrement…

Déterminants : leur, nul, douze, tout, certain, autre, zéro, quelque, tel, cent, neuf, seize, quinze, aucun, soixante, dix, cinquante, vingt, son, second, même, onze …

Conjonctions et prépositions : sans, ni, cependant, mais, que, ou, si, quand, comme, après, sauf, durant, jusque, car, soit, par, or, puisque, malgré, selon, jusqu'ici, avant, sinon…