Discours électoral et discours présidentiel : Une étude lexicale comparative de B. Obama

Jacques Savoy

Institut d'informatique
Université de Neuchâtel – rue Emile Argand 11 - 2009 Neuchâtel - Suisse

Résumé

Cet article présente l'analyse lexicale comparative des discours de B. Obama comme candidat puis comme président. Basé sur un corpus comprenant 113 discours électoraux (janvier à octobre 2008) et 168 discours présidentiels (janvier à juillet 2009), nous constatons que les lemmes les plus fréquents ne se modifient pas avec l'accession du candidat à la présidence. Nous avons tout de même noté un recours significativement plus fréquent à l'adjectif et à l'adverbe dans le discours présidentiel et une abondance plus marquée du pronom dans le discours électoral. La comparaison des mots sur-employés (score Z) permet de dessiner les différences de formes et de thèmes (« McCain », « if », « change » vs. « thank », « reform », « Chrysler »). L'application de diverses formes de lissage permet d'obtenir de meilleures estimations sans modifier significativement les résultats. Afin de mieux cerner les différences entre les deux types de discours, nous avons appliqué des filtres (e.g., élimination des déterminants) avant de déterminer les séquences de bigrammes ou de trigrammes les plus significatives. Au niveau des bigrammes sur-employés, la distinction s'avère assez nette entre le candidat (« Wall Street », « middle class », « tax break ») et le Président (« recovery plan », « new foundation », « important step »). Cette démarcation formelle et thématique se confirme avec les trigrammes (« war in Iraq », « capital gain tax », « world class education » vs. « health care reform », « clean energy economy », « health insurance reform »).

Abstract

This paper describes a lexical comparative study of Obama's speeches as candidate, and as President. Based on a corpus containing 113 electoral speeches (January to October 2008) and 168 presidential speeches (January to July 2009), we found that the most frequent words (or lemmas) do not vary with the arrival of the candidate in the White House. When inspecting the distribution of the Part-of-Speech (POS), we can find that the President uses more frequently adjectives and adverbs while as candidate, B. Obama makes use more often of pronouns. Using the Z score, we were able to define the most significant terms making the difference both in form and content between the two types of speech ("McCain", "if", "change" vs. "thank", "reform", "Chrysler"). Using different smoothing techniques does not modify significantly our results. After applying filters to remove non-pertinent sequences of words, we may define bigrams that best characterizing the electoral ("Wall Street", "middle class", "tax break") or presidential speech ("recovery plan", "new foundation", "important step"). The use of trigrams may confirm the differences between the two types of speech ("war in Iraq", "capital gain tax", "world class education" vs. "health care reform", "clean energy economy", "health insurance reform").

Key-words : discourse analysis, political discourse, lexical comparison, peculiar vocabulary

1. Introduction

La Toile met à notre disposition un nombre considérable de sources d'information et leur facilité d'accès peut nous conduire à être submergé sous un flot de documents. Comment synthétiser un long document ? Comment extraire les mots ou expressions significatifs d'un texte, d'une page Web ou d'un ensemble de discours ? Par exemple, la génération automatique d'un "nuage

de termes" (*term cloud*) a été proposée comme synthèse compacte du contenu d'une page, d'un site Internet ou pour fournir une représentation textuelle à des sources audio ou vidéo (Fuller et al., 2008). Dans le cadre de cette communication, nous souhaitons évaluer empiriquement la *représentation comparative* d'un corpus au moyen d'éléments lexicaux permettant de le différencier d'une autre collection de documents. Contrairement à la génération automatique de résumé, notre démarche s'inscrit dans une perspective comparative, faisant ressortir les caractéristiques propres d'un corpus.

Comme seconde perspective, cet article se situe dans l'analyse des discours politiques et, plus précisément, dans la comparaison lexicale du discours électoral et gouvernemental. Si l'on se souvient que les slogans du candidat Obama tournaient autour des expressions « yes, we can » ou « the change we believe in », est-ce que ces expressions permettent de distinguer le discours du candidat comparé à celui du président ? Bien que notre approche s'appuie sur une analyse lexicale, nous avons également tenu compte d'éléments complémentaires comme la longueur des phrases ou la distribution des parties du discours.

2. Travaux reliés

Dans l'analyse lexicographique et comparative des discours politiques, nous pouvons mentionner les travaux de Labbé and Monière (2003 ; 2008a) qui comparent trois sources de discours gouvernementaux, soit le discours du Trône (Canada), le discours inaugural (Québec) et les déclarations de politique générale (France). Les avantages de cette étude tiennent au fait que cet ensemble de discours est rédigé dans la même langue et couvre une période relativement longue (de 1945 à 2000). On peut également souligner que cette analyse repose sur trois régimes parlementaires. Par contre, le corpus analysé correspond à des discours gouvernementaux qui ne sont pas rédigés dans une perspective électorale. On peut s'attendre à des formulations différentes entre le discours d'un premier ministre en exercice et celui qu'il a tenu pour assurer son élection (Herman, 1974). Selon l'étude de Labbé and Monière (2003), même si les discours gouvernementaux expriment les idées de différents partis politiques, ils ont tendance à être plus similaires que l'on pouvait s'y attendre. Les contraintes institutionnelles ne sont pas étrangères à ce phénomène. Par exemple, la continuité de l'exercice du pouvoir tend à gommer le clivage des partis. Les auteurs soulignent toutefois des modifications temporelles comme la tendance à disposer de discours plus longs au fil des années (plus grande complexité des questions abordées), avec une augmentation sensible de la longueur entre les discours de la IV^{e} et ceux de la V^{e} République.

Afin de comparer deux types de discours, nous devons pouvoir mesurer objectivement la richesse lexicale mais cette dernière notion ne dispose pas d'une définition précise et admise par tous. Nous pouvons tenir compte du nombre de mots, du nombre de mots distincts, du nombre de vocables, de la diversité du vocabulaire ou de sa spécificité, etc. (Baayen, 2008). Pour ce qui concerne le discours gouvernemental, la raison expliquant un accroissement du vocabulaire ne peut pas être attribué à une seule cause clairement définie mais semble survenir avec la prise de pouvoir d'une forte personnalité à l'exemple de P. E. Trudeau au Canada (1968-72) ou, en France, avec M. Rocard (1988) ou P. Bérégovoy (1992).

D'autres travaux en traitement automatique des discours politiques ayant un lien plus ou moins important avec la présente étude peuvent également être mentionnés. Ainsi, on peut s'interroger sur l'identification de l'homme de plume derrière le discours, comme par exemple, T. Sorensen dans l'ombre du Président Kennedy (Carpenter and Seltzer, 1970 ; voir aussi Monière and Labbé, 2006). Nous pourrions également nous appuyer sur une mesure de distance lexicale

(Labbé, 2007) entre deux discours, deux ensembles de discours ou entre quelques leaders politiques (Labbé and Monière, 2003) afin de déterminer leur relative proximité ou éloignement afin de dresser une carte.

3. Notre corpus de discours politiques et traitements préalables

Afin de mener notre étude, nous avons téléchargé [1] les discours tenus par B. Obama comme candidat à la Maison Blanche puis comme président. Ce corpus sera logiquement subdivisé en deux parties. Dans la première, nous retrouvons l'ensemble des discours électoraux du candidat tenus entre le 10 février 2007 (annonce de sa candidature à la présidentielle) et le 31 octobre 2008 (fin officielle de la campagne). Toutefois, nous avons retiré les 37 discours de l'année 2007 afin de fonder notre comparaison uniquement sur les années 2008 (campagne électorale) et 2009 (la présidence). Le corpus électoral comprendra donc 113 discours tandis que le corpus gouvernemental regroupera 168 discours prononcés entre le 20 janvier 2009 (discours d'investiture) et le 23 juillet 2009.

Pour chaque discours, nous avons ajouté quelques informations additionnelles (date, lieu, titre du discours). Ensuite, nous avons remplacé certains codes UTF-8 par leurs équivalents ASCII (e.g., les guillemets " " en «) ainsi que pour les lettres accentuées (e.g., « naïve »). Quelques signes graphiques ont été supprimés. De plus, nous avons remplacé les lettres majuscules par des minuscules sauf si le mot considéré était écrit uniquement avec des majuscules (e.g., « US », « FEMA »). Enfin, nous n'avons pas essayé de normaliser des formes différentes pouvant désigner la même entité comme par exemple avec « US », « United States », « United States of America » ou « USA » (« America », « our country », etc.). Nous estimons que l'auteur (ou les auteurs) conserve la même graphie durant les deux années de notre étude.

Dans Tab. 1, nous avons repris quelques statistiques de notre corpus. Afin d'avoir une idée du volume traité, nous avons indiqué le nombre de mots dans les deux années concernées. Bien que les deux parties renferment un nombre de discours différents (113 et 168), le volume reste assez similaire (294 553 mots vs. 303 273). On peut en déduire que le discours électoral d'Obama s'avère, en moyenne, plus long que l'intervention présidentielle. Dans les deux cas, le mot le plus fréquent demeure le déterminant « the ».

	2008	*2009*
Nombre d'occurrences de mots	294 553	303 273
Mot le plus fréquent	"the" (13 028)	"the" (13 658)
Nombre de formes distinctes	7 663	10 251
Nombre de lemmes	7 925	10 737
Nombre de bigrammes distincts	8 096	12 979
Bigramme le plus fréquent	"health care" (479)	"health care" (585)

Tableau 1 : Quelques statistiques sur les deux parties de notre corpus

Si l'on désire avoir une idée du vocabulaire, nous avons indiqué le nombre de formes distinctes d'une part et, d'autre part, le nombre de lemmes trouvés. Dans le premier cas, toutes les formes fléchies (e.g., « is », « was », « be » ou « soldiers », « soldier ») disposent de leur propre entrée. Sous l'indication « lemme », les formes appartenant à la même entrée dans le dictionnaire sont regroupées (e.g., « be » ou « soldier » dans notre exemple précédent). Par contre, une forme

[1] Depuis les sites officiels soit www.BarackObama.com, respectivement www.WhiteHouse.gov.

peut être ambigüe et être étiquetée selon deux ou plusieurs parties du discours comme le mot « middle » qui peut être analysé comme adjectif ou nom.

Afin de déterminer les parties du discours, nous avons recouru au logiciel d'étiquetage syntaxique automatique de l'Université de Stanford (Toutanova and Manning, 2000), (Toutanova et al., 2003). Ce dernier attribue à chaque mot une étiquette syntaxique et des informations morphologiques dérivées du corpus de Brown (Francis and Kučera, 1982). Par exemple, depuis la phrase « As a nation, we have had our share of debates about the war in Iraq. » le système répondra par «As/IN a/DT nation/NN ,/, we/PRP have/VBP had/VBN our/PRP$ share/NN of/IN debates/NNS about/IN the/DT war/NN in/IN Iraq/NNP ./. ». On y retrouve les étiquettes (Marcus et al., 1993) attachées au nom (NN, nom commun au singulier, NNS nom commun au pluriel, NNP nom propre au singulier), verbe (VB, entrée dans le dictionnaire, VBP présent non 3e personne, VBZ présent 3e personne, VBN participe passé), adjectif (JJ, avec JJR pour le comparatif), pronom personnel (PRP), pronom possessif (PRP$), préposition (IN), adverbe (RB). Depuis chaque mot accompagné de son analyse morphologique, nous pouvons retrouver l'entrée dans le dictionnaire, essentiellement pour les noms par suppression du pluriel (e.g., « jobs/NNS » → « job/NN ») et par substitution de la forme fléchie des verbes (e.g., « argues/VBZ » → « argue/VB »).

Ce traitement automatique n'est pas exempt d'erreurs ou de syntagmes dont l'étiquetage proposé reste sujet à discussion, comme par exemple pour le groupe nominal « Senate Foreign Relations Committee ». Une première solution consiste à donner la même partie du discours aux quatre éléments (soit nom propre ou « Senate/NNP Foreign/NNP Relations/NNP Committee/NNP »). Comme alternative, on attribue l'étiquette nom propre au pluriel au mot « Relations/NNPS ».

Les analyses subséquentes pourront donc se focaliser sur les mots (ou formes comme « class » et « classes ») ou les lemmes (« class/NN »). Cependant, nous pouvons également recourir à des entités sémantiquement plus précises comme une séquence de deux mots (bigramme) ou de trois mots (trigramme). Toutefois, l'étude de telles séquences ne révèle pas toujours une sémantique très appropriée (e.g., « of the » ou « I will » correspond à des bigrammes très fréquents). Nous avons donc retenu les séquences répondant à certaines contraintes. Ainsi les bigrammes doivent correspondre au format « nom nom » ou « adjectif nom » (e.g., « interest rate » ou « Attorney General »). Pour les trigrammes les constructions autorisées sont : « nom nom nom », « nom adjectif nom », « adjectif nom nom », « adjectif adjectif nom » ou « nom préposition nom » (e.g., « sense of fairness » ou « long term challenge »). Notre étude pouvant choisir différents éléments lexicaux, nous utiliserons le mot *terme* de manière générique pour couvrir tant les mots, les lemmes ou des séquences de bigrammes ou trigrammes.

4. Méthodologie d'analyse comparative des discours

Afin de comparer deux corpus, nous pouvons nous appuyer, dans une première étape, sur la richesse du vocabulaire (section 4.1) ou la longueur des phrases et la distribution des parties du discours (section 4.2). Afin d'extraire le vocabulaire caractéristique d'un corpus par rapport à une référence, nous nous sommes appuyés sur l'approche proposée par Muller (1992) (score Z, section 4.3). Toutefois, nous pouvons affiner l'estimation des probabilités d'occurrence en recourant à diverses techniques de lissage (section 4.4).

4.1. Richesse du vocabulaire

Un regard attentif sur Tab. 1 indique que, bien que les deux corpus renferment un volume similaire de mots, le discours présidentiel présente un nombre plus élevé de formes distinctes, de lemmes et de bigrammes. Cet accroissement signale un vocabulaire plus riche chez le

Président, abordant des thématiques plus variées. Pour le candidat, il existe la nécessité de tenir un discours simple et d'insister sur thèmes essentiels qu'il désire aborder.

Afin de vérifier cette hypothèse, nous pourrions prédire le nombre attendu de lemmes dans le discours présidentiel sur la base du discours électoral. Sur ce dernier corpus, nous avons modélisé au moyen de la régression linaire la relation entre le logarithme du nombre de mots et le logarithme du nombre de lemmes. Pour estimer les paramètres de cette droite, nous avons compté le nombre de lemmes par tranche de 1000 mots dans le corpus électoral ($R^2 = 0{,}9782$). Sur ce modèle statistique (calcul effectué à l'aide du logiciel R; Crawley, 2007), le nombre de lemmes prévu pour 303 273 mots (taille du discours présidentiel) s'élève à 8 913 (avec un intervalle de confiance à 95 % de 7 726,5 à 10 282). La valeur observée (10 737) se situe en dehors de cet intervalle, indiquant clairement une richesse lexicale plus importante chez le Président.

4.2. Comparaison des moyennes des longueurs de phrase

Au niveau des phrases, on remarque également des différences. Le candidat possède une formulation plus allongée avec un nombre moyen de mots de 21,33 par phrase (médiane 19; écart-type 13,46), tandis que la phrase typique du discours gouvernemental est plus courte (moyenne 17,96; médiane 16; écart-type 14,5). Cette différence entre les deux moyennes s'avère significative ($\alpha = 0{,}01$, *t-test)*.

Partie	*Observé*		*Total*	*Attendu*	
	Candidat	*Président*		*Candidat*	*Président*
Nom	72 174	74 617	146 791	72 096	74 695
Verbe	54 253	56 717	110 970	54 502	56 468
Adjectif	18 511	20 488	38 999	19 154	19 845
Adverbe	17 064	18 347	35 411	17 392	18 019
Pronom	30 742	29 080	59 822	29 381	30 441
Autres	103 852	108 042	211 822	104 071	107 823
Total	296 596	307 291	603 887	296 596	307 291

Tableau 2 : Distribution des catégories syntaxiques dans les discours d'Obama candidat et ceux du Président Obama (nombre observé et attendu)

L'emploi des diverses parties du discours entre les corpus permet de compléter cette première analyse. Dans Tab. 2, sous les colonnes « Observé », nous avons indiqué le nombre de noms, verbes, adjectifs, adverbes et pronoms apparaissant dans les discours électoraux et présidentiels. Dans une dernière classe nommée « Autres », nous avons regroupé toutes les autres catégories (préposition, auxiliaire, nombre, etc.) ainsi que les signes de ponctuation. La dernière ligne et la quatrième colonne indiquent les différents totaux. Enfin, les colonnes sous l'étiquette « Attendu » indiquent le nombre attendu de chaque catégorie syntaxique sous l'hypothèse que la même distribution est sous-jacente aux deux types de discours.

En appliquant un test du χ^2 sur la base de ces observations, on constate que les deux distributions sont statistiquement différentes ($\chi^2 = 181{,}7$; $\alpha = 0{,}01$). Au niveau du volume des noms et des verbes, les discours électoral et présidentiel ne se différencient pas. Par contre, on remarque que le candidat utilise de manière nettement moins fréquente les adjectifs et quelque peu les adverbes. Par contre le candidat recourait plus fréquemment aux pronoms (e.g., « we » et « I »). Dans ce cas, on notera un sur-emploi du « je » que l'on retrouve également lors de la dernière élection présidentielle française (Labbé and Monière, 2008b) (voir également Tab. 3).

4.3. Termes sur- et sous-employés

Afin de comparer nos deux corpus, nous pourrions prendre en compte les lemmes les plus fréquents repris dans Tab. 3. Ces informations ne permettent pas de distinguer clairement les écrits du candidat de ceux du président. On constate seulement deux permutations entre les deux corpus, soit entre les lemmes « we » et « to » d'une part et, d'autre part, entre « I » et « in ». On peut y voir un indice de la présence du même auteur derrière les deux types de discours (Baayen et al. 2002). A titre de comparaison, nous avons également indiqué les dix lemmes les plus fréquents dans les 71 discours électoraux prononcés par le Sénateur J. McCain en 2008. Avec cette information complémentaire, on constate immédiatement la similitude entre ces deux types de discours prononcés par B. Obama.

Rang	*Obama - Candidat*		*Obama - Président*		*McCain - Candidat*	
	Fréquence	*Lemme*	*Fréquence*	*Lemme*	*Fréquence*	*Lemme*
1	13 028	the / DT	13 658	the / DT	7 759	the / DT
2	11 695	be / VB	13 279	be / VB	6 157	and / CC
3	10 951	and / CC	11 387	and / CC	5 413	to / TO
4	10 472	we / PRP	10 694	to / TO	5 174	be / VB
5	9 071	to / TO	10 671	we / PRP	4 773	of / IN
6	6 985	of / IN	8 238	of / IN	4 199	we / PRP
7	6 344	an / DT	6 891	an / DT	3 480	I / PRP
8	5 596	I / PRP	5 351	in / IN	3 344	an / DT
9	5 286	in / IN	4 814	I / PRP	3 125	in / IN
10	4 072	have / VB	4 530	have / VB	2 048	have / VB
Total	296 596		307 291		155 097	

Tableau 3 : Les dix lemmes les plus fréquents chez le candidat Obama (gauche), le Président Obama (centre) et le candidat McCain (droite)

Une meilleure approche consiste à déterminer les termes (mot, lemme, bigramme) les plus représentatifs d'un corpus (noté C_1) en les comparant à une référence (corpus noté C_2). Comme l'illustre Tab. 4, on peut répartir le nombre d'occurrences d'un terme donné ω entre le corpus C_1 (valeur a) et C_2 (valeur b). Sur l'ensemble des documents retenus, le mot ω considéré apparaît $a+b$ fois. Nous pouvons alors estimer la probabilité d'occurrence du terme ω dans le corpus C par $(a+b)/n$ en utilisant le principe du maximum de vraisemblance (MLE).

	C_1	C_2	$C = C_1 \cup C_2$
ω	a	b	$a + b$
pas ω	c	d	$c + d$
	$a + c$	$b + d$	$n = a + b + c + d$

Tableau 4 : Exemple d'une table de contingence pour le terme ω

Muller (1992) suggère d'utiliser cette estimation pour calculer un score Z normalisé que l'on associe à chaque terme selon la formule 1. Dans ce cas, on admet que la distribution d'occurrence du terme ω suit une loi binomiale avec comme paramètres p (probabilité de succès) et n (le nombre de répétition). Dans notre cas, le nombre attendu d'occurrences du terme ω dans le corpus C_1 s'élève à $p \cdot (a+c) = ((a+b)/n) \cdot (a+c)$, avec une variance de $p \cdot (1-p) \cdot (a+c)$. Le score Z du terme ω correspond donc à la différence entre le nombre observé (a) et le nombre attendu

(p·(a+c)) divisé par la racine carrée de la variance (le score Z se modélise comme une variable aléatoire normale centrée et réduite).

$$\text{score Z} = \frac{a-(p\cdot(a+c))}{\sqrt{p\cdot(1-p)\cdot(a+c)}} \quad (1)$$

Avec cette formulation, nous pouvons définir les sur-emplois dans un corpus C_1 par rapport au corpus de référence comme les mots ayant un score Z positif et supérieur à un seuil δ donné (e.g., δ = 2). De façon similaire, on définit les sous-emplois par une valeur Z négative et inférieure à un seuil fixé (e.g., -δ). Par exemple, on compte quatre occurrences du mot « recovery » dans les discours du candidat Obama [2] mais 231 fois dans les discours présidentiels (voir Tab. 5). Nous pouvons estimer sa probabilité d'apparition comme p = 235 / 597826 = 0,0003931. Dans le cadre du corpus présidentiel, nous pouvons attendre 0,0003931 . 303273 = 118,5 occurrences de ce terme. Le score Z indiqué par la formule 2 s'avère positif et très élevé (10,24), indiquant clairement un sur-emploi du mot « recovery » dans le discours présidentiel par rapport au discours électoral.

$$\text{score Z(recovery)} = \frac{a-(p\cdot(a+c))}{\sqrt{p\cdot(1-p)\cdot(a+c)}} = \frac{231-(0,000393\cdot 303273)}{\sqrt{0,000393\cdot(1-0,000393)\cdot 303273}} = 10,24 \quad (2)$$

	Président	*Candidat*	$C = C_1 \cup C_2$
"recovery"	231	4	235
autres	303 042	294 549	597 591
	303 273	294 553	597 826

Tableau 5 : Table de contingence pour le mot "recovery"

La limite de δ = 2 que nous avons adoptée demeure un peu arbitraire et correspond, pour une loi normale centrée et réduite, à un ensemble de valeurs couvrant 2,28 % des observations. Si l'on analyse les mots, nous retrouvons 444 mots sur-employés dans le discours électoral et 569 dans le discours présidentiel, pour un total de 10 967 mots. Les sur-emplois couvrent donc respectivement 4,05 % des observations chez le candidat et 5,19 % chez le président. Au niveau des bigrammes, nous avons 18 876 termes dont 289 (ou 1,53 %) sont surreprésentés dans le discours électoral et 106 (ou 0,56 %) dans le discours présidentiel.

4.4. Correction « LRNE » des résultats d'emploi des termes

Définir un terme comme sur-employé se fonde sur une estimation de sa probabilité d'occurrence. Cette dernière repose habituellement sur le maximum de vraisemblance (MLE) qui nous conduit à estimer *p comme (a+b)/n*. Cette méthode possède toutefois quelques lacunes. Par exemple, l'estimation d'un terme qui n'est jamais apparu dans le corpus donne la valeur 0. Or, il est reconnu que la distribution des mots suit une distribution de type LNRE (*Large Number of Rare Events*; Baayen, 2001). Nous ne pouvons donc pas négliger la classe des termes encore inconnus. Comme nous savons que l'ensemble des *hapax (terme de fréquence unitaire) couvre habituellement une grande proportion des mots d'un corpus, il s'avère d'autant plus surprenant d'exclure la possibilité d'apparition d'un nouveau mot.*

[2] Ces quatre occurrences apparaissent dans quatre discours différents (21 septembre, 2, 3 et 9 octobre).

Dans cette perspective, nous pouvons lisser les estimations des probabilités (Manning and Schütze, 2000). Une première approche simple consiste à ajouter une unité au numérateur de notre estimation et d'ajouter, en complément, au dénominateur la taille du vocabulaire retenu. Cette formulation se généralise (loi de Lidstone) en lissant toute probabilité par la formule $p = (a+b+\lambda) / (n+\lambda.|V|)$, avec λ un paramètre de lissage et $|V|$ indiquant la taille de notre vocabulaire (e.g., 10 967 mots, 12 940 lemmes ou 18 876 bigrammes).

Afin de fixer la valeur λ, plusieurs choix s'avèrent possibles comme $\lambda = 1$ (Laplace), $\lambda = 0{,}5$ (ELE ou *Expected Likelihood Estimation*) ou sélectionner une valeur très petite comme, par exemple, à $\lambda = 1/(b+d)$. Ces diverses solutions réduisent, plus ou moins selon la valeur attribuée à λ, les estimations et donc la probabilité associée à l'occurrence des termes. Toutefois, nous ne disposons pas de théorie justifiant un choix précis pour le paramètre λ. De plus, cette stratégie donne parfois des résultats moins précis que le maximum de vraisemblance, en particulier pour les mots apparaissant dans l'échantillon (Gale and Church, 1994). Par contre, l'implémentation de cette approche demeure simple.

Comme alternative, nous avons implémenté et testé une approche dérivée de la famille des techniques de Good-Turing (noté GT) (Sampson, 2001). Dans ce cas, l'estimation associée à l'ensemble des termes inconnus dépend de la fréquence des *hapax*. Si cette dernière valeur est élevée, l'apparition de nouveaux termes sera d'autant plus probable et, par conséquent, la probabilité associée aux nouveaux termes devra être plus forte. En limitant notre analyse aux valeurs Z les plus fortes, les scores Z restent assez similaires que l'on recourt à une estimation de type MLE ou à l'une des méthodes de lissage.

Par exemple, si l'on inspecte les bigrammes possédant un score Z supérieur à deux ($\delta = 2$), on retrouve 252 observations selon l'estimation MLE, 324 avec le lissage Lidstone ($\lambda = 0{,}3$) ou 403 avec le lissage GT. Avec une limite $\delta = 3$, les différences s'amenuisent nettement avec 74 cas selon l'estimation MLE et 78 avec le lissage GT. En complément, nous pouvons signaler que parmi l'ensemble des cinquante scores les plus élevés, seulement quatre termes divergent entre le lissage GT et l'estimation MLE (e.g., le bigramme « common sense » possède le rang 45 dans un cas, 60 dans l'autre). La même analyse indique qu'entre le lissage $\lambda = 0{,}3$ et l'estimation MLE, nous avons trois bigrammes différents.

5. Analyse lexicale et comparative du discours électoral et gouvernemental

En nous limitant à la dernière élection présidentielle américaine et au seul candidat vainqueur de cette élection, nous pouvons extraire les mots caractéristiques du discours électoral et les comparer à ceux du discours présidentiel (section 5.1). Souvent la sémantique associée aux mots reste trop ambigüe et le recours à des séquences de deux mots (section 5.2) ou trois mots permet d'éclairer plus précisément les différences. Toutefois l'application de filtres permet de mieux sélectionner les séquences pertinentes à analyser.

5.1. Analyse des mots isolés

Comme premier niveau d'analyse comparative entre les discours électoraux et présidentiels, nous avons retenus les mots isolés. Pour les deux parties, nous avons calculé le score Z (lissage $\lambda = 0{,}3$) et nous avons tenu compte essentiellement des mots présentant les vingt valeurs les plus élevées. Dans le Tab. 6 nous avons indiqué les mots les plus représentatifs accompagnés d'une part de leur score Z et, d'autre part, de leur fréquence d'occurrence.

Obama - Candidat			*Obama - Président*		
Score Z	*Fréquence*	*Forme*	*Score Z*	*Fréquence*	*Forme*
18,23	696	McCain	12,96	611	thank
15,08	958	tax	10,73	241	everybody
13,14	489	Street	10,36	231	recovery
12,15	472	Senator	9,89	457	reform
12,12	321	Bush	8,10	177	extraordinary
10,88	347	election	7,96	1 406	so
10,64	345	Wall	7,86	653	going
10,08	684	change	7,41	834	these
9,96	553	Washington	7,28	328	very
9,68	3 575	not	7,10	10 694	to
9,60	739	President	6,95	3 376	are
9,29	937	he	6,79	1 889	as
8,85	309	John	6,67	1 550	all
8,71	272	policy	6,49	95	team
8,58	1063	if	6,42	267	effort
8,38	282	campaign	6,35	231	budget
8,34	1 047	need	6,34	234	already
8,33	326	class	6,32	138	foundation
8,25	1 170	when	6,20	90	Chrysler
8,20	309	middle	6,14	8 238	of

Tableau 6 : Les vingt mots sur-employés par le candidat Obama (gauche) et le Président Obama (droite)

Toute analyse de discours doit tenir compte des spécificités imposées par la forme ou le contexte. Ainsi, dans son discours électoral, le candidat Obama se présente en opposition au Sénateur John McCain et les sur-emplois soulignent cette caractéristique (« election », « President », « campaign », « McCain », « Senator », « John »). Il fait référence au passé (« Bush », « Washington ») et indique que les politiques doivent changer s'il est élu (« change », "if »). Mais la sémantique devient plus floue et incertaine si l'on considère d'autres mots sur-employés. Dans ce cas, on retrouve les mots « Street », « Wall » ou « class », « middle », « policy » et « tax ». On peut certes inférer des associations plausibles (« Wall Street » ou « middle class ») sans que la sémantique gagne vraiment en acuité.

De son côté, le discours présidentiel est également sujet à des contraintes de formes. Le président remercie l'assemblée ou les personnalités présentes (« thank », « everybody »). Quelques mots permettent de voir se dessiner les problèmes importants de ce début de mandat (« reform », « recovey », « budget », « foundation »), voire les questions plus pointues (« Chrysler »). Dans ce cas également, le sur-emploi de diverses formes n'a pas d'explication immédiate (e.g., « so », « these », « very », « are »).

5.2. Analyse des bigrammes et trigrammes

Afin d'améliorer la représentativité, nous pouvons recourir à des séquences de deux mots (bigrammes) après filtrage (e.g., élimination des prépositions ou déterminants). Dans Tab. 7, nous avons extrait les vingt bigrammes ayant le score Z le plus élevé et, en complément, leur fréquence d'occurrences. Nous avons procédé de la même manière avec les trigrammes dont les vingt plus représentatifs sont indiqués dans Tab. 8. Parfois, les entrées de ces deux tableaux peuvent se compléter. Ainsi, si dans le discours électoral le bigramme "capital gain" reste ambigu, les trigrammes en précise le sens (« capital gain tax »).

Obama - Candidat			*Obama - Président*		
Score Z	*Fréquence*	*Bigramme*	*Score Z*	*Fréquence*	*Bigramme*
15,29	384	Senator McCain	6,76	98	care reform
13,01	289	John McCain	5,76	67	recovery plan
12,05	300	Wall Street	5,48	118	clean energy
9,80	245	middle class	5,45	60	recovery act
9,52	148	George Bush	4,97	50	new foundation
7,98	127	Main Street	4,61	43	big round
7,71	131	tax break	4,09	42	Prime Minister
7,64	214	tax cut	3,85	33	economic recovery
7,07	83	oil company	3,78	37	community college
6,87	76	more year	3,72	31	New Jersey
6,35	65	rescue plan	3,51	25	important step
5,59	53	real change	3,38	40	economic growth
5,40	129	new job	3,35	103	care system
5,36	54	gain tax	3,21	21	American recovery
5,34	56	capital gain	3,18	26	higher education
5,30	48	income tax	3,13	20	Jon Corzine
5,21	49	same kind	3,10	25	good morning
5,18	46	more support	2,99	26	North Korea
5,07	132	insurance company	2,97	18	President Medvedev
5,01	48	gas price	2,96	30	same time

Tableau 7 : Les vingt bigrammes sur-employés par le candidat Obama (gauche) et le président Obama (droite)

Obama - Candidat			*Obama - Président*		
Score Z	*Fréq.*	*Trigramme*	*Score Z*	*Fréq.*	*Trigramme*
6,27	95	war in Iraq	6,98	98	health care reform
5,57	54	capital gain tax	5,32	54	round of applause
5,09	39	common sense regulation	3,56	103	health care system
5,01	45	world class education	3,48	26	clean energy economy
4,46	30	Bush tax cut	2,88	16	house of representative
4,46	49	middle class family	2,88	16	health insurance reform
4,46	61	source of energy	2,69	14	proportion of college
4,37	53	last eight year	2,61	30	rule of law
4,15	26	month in Iraq	2,59	13	next 10 year
4,00	29	next four year	2,52	15	quality affordable health
3,99	24	other's success	2,48	12	lot of work
3,74	32	middle class tax	2,48	12	health care spending
3,63	20	uncertainty for America	2,48	12	health care coverage
3,63	20	kind of change	2,48	12	good morning everybody
3,59	22	kind of health	2,48	12	comprehensive health care
3,54	19	new green job	2,42	14	kind of energy
3,51	38	early childhood education	2,38	11	good afternoon everybody
3,50	21	side of Chicago	2,38	11	Foundation for growth
3,47	25	class tax cut	2,26	10	stem cell research
3,45	18	vote on November	2,26	10	piece of legislation
3,45	18	affordable accessible health	2,19	12	range of issue

Tableau 8 : Les vingt trigrammes sur-employés par Obama candidat (gauche) et Obama Président (droite)

Comme pour les mots, le contexte impose ses formes au discours présidentiel (« Prime Minister », « good morning », « big round (of applause) »), mais également l'actualité du moment (Jon Corzine, ancien gouverneur du New Jersey, et candidat démocrate malheureux à ce poste lors de l'élection du 3 novembre 2009). Mais les thèmes caractéristiques du président, selon notre analyse lexicale, ressortent plus clairement telles que les réformes (« health care reform », « clean energy ») la nécessité d'un plan de sauvetage de l'économie (« recovery plan », « economic recovery ») ou la volonté d'un changement profond (« clean energy economy », « new foundation »). De plus, le président utilise plus le terme « economic » et il se doit de parler du « budget ».

Obama candidat se distingue du Président par le recours à des formes générales pour parler de l'économie réelle (« Main Street ») du monde de la finance (« Wall Street »), la volonté de créer de nouveaux emplois (« new job(s) », « new green job(s) »), la réduction des impôts (« middle class tax », « tax cut ») ou leur augmentation dans d'autre cas (« capital gain tax »), la santé (« affordable accessible health »), l'éducation (« world class education ») ou sa préoccupation concernant la guerre en Iraq (« war in Iraq », « month(s) in Iraq »). Le Président réduira les thèmes récurrents abordés et sera plus précis (e.g., « health care reform », « health care spending », « health care coverage », « quality affordable health »). Le thème des impôts (réduction « tax cut ») ou création (« capital gain tax ») n'est pas dans l'agenda présidentiel.

6. Conclusion

La comparaison lexicale du discours électoral et gouvernemental que nous avons proposée reposait essentiellement sur trois approches, à savoir les mots isolés et les lemmes, les bigrammes et finalement les trigrammes. Dans les deux derniers cas, nous avons proposé d'appliquer des filtres afin d'éliminer des séquences ne présentant *a priori* qu'un intérêt marginal. De plus, leur usage impose la présence d'un corpus plus conséquent afin d'éviter de devoir analyser des séquences de fréquences très faibles (e.g., entre trois et cinq occurrences). Pour définir une représentation compacte, nous avons retenu les vingt termes ayant le score Z le plus élevé. Sur cet ensemble, le lissage des probabilités soit par la loi de Lidstone, soit par la technique de Good-Turing, ne modifie pas significativement les éléments retenus.

Au niveau de la représentation, les mots isolés n'apportent que peu d'indice sémantique. Certes, nous pouvons parfois rencontrer une entité clairement identifiée (« Chrysler ») ou l'indication de l'importance des pronoms personnels (« I », « we ») dans le discours électoral. Les bigrammes (« oil company ») et les trigrammes (« stem cell research ») permettent de mieux cerner la portée sémantique et l'intention de l'auteur. Afin de générer automatiquement une représentation compacte (« term cloud »), nous devons encore pouvoir fusionner de manière appropriée la liste des mots (et/ou lemmes), celle des bigrammes et celle des trigrammes. Comme autre perspective, nous pourrions recourir à ces informations pour classifier automatiquement un discours comme celui du candidat ou du Président (Savoy and Zubareyva, 2010).

Remerciements

Cette recherche a été financée par le Fonds national suisse pour la recherche scientifique (subside n^0 200021-124389).

Références

Baayen H.R. (2001). *Word Frequency Distributions*. Dordrecht : Kluwer Academic Press.

Baayen H.R. (2008). *Analysis Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Baayen H.R., van Halteren H., Neijt A. and Tweedie F. (2002). An experiment in authorship attribution. In *Actes JADT 2002*, St Malo.

Carpenter R.H. and Seltzer R.V. (1970). On Nixon's Kennedy style. *Speaker and Gavel*, 7(41) : 41-43.

Crawley M.J. (2007). *The **R** Book*. Chichester : John Wiley & Sons.

Francis W.N. and Kučera H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Fuller M., Tsagkias M., Newman E., Besser J., Larson M., Jones G.J.F. and de Rijke M. (2008). Using term clouds to represent segment-level semantic content of podcasts. In *Proceedings 2nd SIGIR Workshop on Searching Conversational Speech*, Singapore.

Gale W.A. and Church K.W. (1994). What is wrong with adding one? In Oostdijk, N. and de Hann, P., editors, *Corpus-Based Research into Language*, Amsterdam-Atlanta: Rodopi.

Herman V. (1974). What governments say and what governments do: An analysis of post-war Queen's speeches. *Parliamentary Affairs*, 28(1) : 22-31.

Kilgarriff A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1) : 97-133.

Labbé D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics*, 14(1) : 33-80.

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

Labbé D. and Monière D. (2008a). *Les mots qui nous gouvernent. Le discours des premiers ministres québécois: 1960-2005*. Montréal : Monière-Wollank.

Labbé D. and Monière D. (2008b). Je est-il un autre ? In *Actes JADT 2008*, Lyon, pp. 647-656.

Manning C.D. and Schütze H. (2000). *Foundations of Statistical Natural Language Processing*. Cambridge : The MIT Press.

Marcus M.P., Santorini B. and Marcinkiewicz M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2) : 313-330.

Monière D. and Labbé D. (2006) L'influence des plumes de l'ombre sur les discours des politiciens. In *Actes JADT 2006*, Besançon, pp. 687-696.

Muller C. (1992). *Principes et méthodes de statistique lexicale*. Paris : Honoré Champion.

Nugues P.M. (2006). *An Introduction to Language Processing with Perl and Prolog*. Berlin : Springer-Verlag.

Sampson G. (2001). *Empirical Linguistics*. London : Continuum.

Savoy J. and Zubaryeva O. (20010). Classification automatique d'opinions dans la blogosphère In *Actes JADT 2010*, Rome.

Toutanova K. and Manning C. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagging. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pp. 63-70.

Toutanova K., Klein D., Manning C. and Singer Y. (2003). Feature-rich part-of-speech tagging with a cyclid dependency network. In *Proceedings of HLT-NAACL 2003*, pp. 252-259.