Représentation comparative : Application au discours électoral en Suisse, France et Etats-Unis

Jacques Savoy

Institut d'informatique
Université de Neuchâtel, rue Emile Argand 11, 2009 Neuchâtel (Suisse)
Jacques.Savoy@unine.ch

RÉSUMÉ. *Dans cet article, nous décrivons et analysons plusieurs méthodes afin d'extraire les mots ou expressions les plus représentatifs d'un corpus en comparaison avec un corpus de référence. L'emploi de la fréquence d'occurrence ou le rang des termes les plus fréquents peut fournir une première synthèse. Une meilleure approche consiste à s'appuyer sur l'hypothèse d'une distribution binomiale des mots et le calcul d'un score normalisé (score Z) mettant en lumière les termes ou expressions comparativement les plus usités dans le corpus. Afin d'illustrer la qualité de la représentation ainsi obtenue, nous avons appliqué notre stratégie sur des discours électoraux suisses, français ou américains.*

ABSTRACT. *This paper describes some methods to automatically extract terms or sequences of terms closely reflecting the content of a corpus or a Web site by comparison of a given corpus. The frequency of occurrences or the rank of the most frequent terms may provide a first overview. The suggested method is based on the terms distribution according to a binomial process and we proposed to compute a normalized Z-score to define the most appropriate terms within a comparative perspective. Examples based on Swiss, French and US political speeches show the usefulness of the suggested method.*

MOTS-CLÉS : *Résumé automatique, indexation, analyse du discours, discours politique, distribution lexicale.*

KEYWORDS: *Summarization, indexing, discourse analysis, political speeches, probabilistic word distribution.*

1. Introduction

Internet, outil de communication et de partage du savoir, continue à progresser en offrant ses services à un nombre croissant de personnes. Celles-ci ne se contentent plus d'être de simples consommateurs d'information gravitant autour des moteurs de recherche (Witten *et al.*, 2007), consultant des horaires ou la météo,

réservant des chambres d'hôtel, ou achetant livres et CDs de musique. Les internautes occupent également un rôle de producteur d'information en participant à des forums de discussion, en rédigeant leurs journaux ou en répondant à des billets placés dans des *blogs* ou autres sites de réseaux sociaux.

Afin de représenter, de synthétiser ou de guider les utilisateurs dans un volume croissant d'information, de nombreuses perspectives d'application s'ouvrent, domaines que l'on peut regrouper sous le terme général de "fouille de textes" (Konchady, 2006). Dans cette perspective, nous nous sommes intéressé à la mise au point d'outils automatiques permettant d'extraire les termes (mot isolé, bigramme ou trigramme) les plus caractéristiques d'un site Internet ou d'un corpus. Comme en recherche d'information (Boughanem *et al*., 2008), notre objectif consiste à indexer et à représenter d'une manière compacte une page, un site ou un ensemble de documents formant un corpus. Dans notre contexte, une telle représentation doit mettre en évidence le contenu sémantique d'un site en comparaison avec d'autres sites voire du même site à une date antérieure.

Disposant d'une telle représentation, nous pourrions répondre à divers besoins comme le souci des entreprises de suivre en continu l'évolution de leurs concurrents via leurs sites web (veille technologique) ou le suivi d'événements sociaux ou politiques (TDT, *Topic Detection & Tracking*). La *blogosphère* (Fautsch *et al*., 2008) présente également un intérêt afin de suivre dans une perspective comparative les sentiments des internautes, exprimés dans des billets d'information ou, dans une optique dynamique, en mettant en lumière l'évolution temporelle des intérêts ou sentiments.

Cet article explore quelques méthodes afin de définir et d'analyser diverses stratégies de représentation comparative. Nous nous limiterons cependant à des informations de nature textuelle pouvant correspondre à un site Internet ou à un corpus. Nos exemples seront extraits de discours électoraux suisses, français ou américains. Dans cette perspective, quelques travaux reliés à la génération automatique de résumés ou à l'extraction de termes significatifs seront présentés dans la deuxième section. Cette section sera complétée par un survol des travaux antérieurs touchant l'analyse des discours politiques. Les corpus servant d'exemples à nos propos seront décrits dans la troisième section. Les diverses stratégies d'extraction seront décrites dans la quatrième section. Enfin, en recourant à nos outils, une comparaison des discours électoraux en Suisse (élection fédérale d'octobre 2007) et une comparaison franco-suisse avec l'élection présidentielle de 2007 seront présentées dans la dernière section. Une analyse de la dernière campagne présidentielle américaine (2008) complétera cette ultime partie.

2. Résumé automatique et analyse de discours politiques

Cette étude repose sur deux aspects complémentaires. En premier la mise au point et l'analyse de diverses approches afin de représenter de manière plus ou moins compacte un corpus. Contrairement à plusieurs approches suggérées par d'autres

auteurs, notre stratégie repose sur le besoin de fournir une représentation mettant clairement en lumière les différences entre plusieurs corpus. Les travaux reliés à cette première thématique sont regroupés dans la section 2.1. Comme nous avons appliqué et analysé les méthodes suggérées dans le cadre des discours électoraux, la section 2.2 présente un survol des travaux antérieurs dans ce domaine d'application.

2.1. Résumé automatique

Dans la génération automatique de résumés (Mani *et al*., 1999), la phrase constitue la structure fondamentale la plus souvent retenue. En effet, il s'avère souvent trop difficile de comprendre, d'interpréter puis de générer un résumé sur la base de fractions de phrases que le système devra ensuite lier tout en garantissant une bonne lisibilité. Le choix de la phrase permet également de supprimer, partiellement pour le moins, les difficultés liées aux coréférences (Strube *et al*., 1996), (Nugues, 2006) (par exemple, références anaphoriques et pronominales). Bien que des travaux récents recourant à des méthodes sophistiquées aient permis de faire quelques progrès dans cette direction, la génération automatique de résumé peut être vue essentiellement comme un problème d'extraction des phrases les plus significatives.

Dans cette perspective, Goldstein *et al*. (1999) distinguent entre deux types de résumés, à savoir le résumé générique ou en réponse à une requête. Cette distinction s'avère pertinente dans le choix des phrases à extraire du document. Ainsi on sélectionnera soit celles qui décrivent le mieux le contenu proprement-dit ou celles qui répondent le mieux à la requête. D'autres critères de choix peuvent s'ajouter comme la longueur et le style de la phrase mais la sélection s'opère essentiellement sur la présence et la pondération des termes contenus dans la phrase. On considère généralement que la fréquence d'occurrence (ou fréquence lexicale notée *tf*) et l'inverse de la fréquence documentaire (*idf*) constituent des facteurs déterminants.

En ce qui concerne l'efficacité de la pondération *tf idf*, les expériences menées n'ont toutefois pas abouti à des conclusions toujours concordantes (Paice, 1990), (Orasan *et al*., 2004). Ainsi, parmi les autres critères intéressants, on pourrait retenir la position de la phrase et sa longueur, deux caractéristiques qui semblent influencer la qualité du résultat final (Kupiec *et al*., 1995). Ces auteurs ajoutent que les groupes nominaux fréquents devraient bénéficier d'un avantage, de même que les termes du titre du document, les mots écrits en majuscules ou les entités nommées.

Parfois le flux d'information n'est pas vraiment cohérent et la structure du discours (ou du document) demeure lacunaire, rendant l'extraction de phrases complètes plus ardue et générant souvent un résumé peu cohérent. Dans ce but, Berger & Mittal (2000) proposent de déterminer d'abord les mots à inclure dans un résumé selon leur fréquence d'occurrence (*tf*) ou selon leur probabilité d'apparition prédite par un modèle de langue (ce qui requiert toutefois un apprentissage). Ensuite, l'ordre des mots dans le résumé final doit être établi en fonction de séquences similaires trouvées dans le (ou les) document(s). Une telle approche peut

être appliquée sur des pages *web* ou la *blogosphère* marquée par l'absence de structure argumentative précise.

De manière plus simple, nous pouvons ignorer l'ordre des termes et limiter la représentation à une liste de mots isolés. L'expression "nuage de termes" (*term cloud*) a été suggérée pour une telle représentation décrivant de manière compacte le contenu d'une page (la figure 1 présente un tel exemple qui se fonde uniquement sur la fréquence d'occurrence[1]). L'importance relative de chaque élément s'illustre par des modifications de la taille de la fonte ou de l'ancre (gras, italique) voire de la couleur, de la police ou de l'emplacement (ou l'ordre). Une telle représentation peut également servir à illustrer le contenu global ou les divers passages d'une séquence audio ou vidéo (Fuller *et al*., 2008).

***Figure 1** : Exemple d'un nuage de termes décrivant l'article que vous lisez (basé sur la fréquence lexicale des vocables)*

2.2. Analyse des discours politiques

Dans l'analyse lexicographique et comparative des discours politiques, nous pouvons mentionner les travaux de Labbé & Monière (2003; 2008) 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. En revanche, 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

[1] La figure 1 a été obtenue grâce au logiciel disponible à l'adresse www.wordle.net

de (Labbé *et al*., 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 (présence de la télévision, 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 *et al*., 1970) (voir également (Monière *et al*., 2006), (Baayen *et al*., 2002)). 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 *et al*., 2003) afin de déterminer leur relative proximité ou leur éloignement pour dresser une carte selon leur affinité et leur proximité lexicale.

3. Corpus d'évaluation, acquisition et prétraitement

Afin de représenter de manière *comparative* un corpus et de juger de la qualité du résultat obtenu, il serait souhaitable de disposer d'une collection test et d'une métrique adaptée. Ces deux derniers éléments n'étant pas disponibles et dans un souci de présenter une justification expérimentale, nous avons décidé d'appliquer les méthodes d'extraction proposées et de les comparer sur la base de divers corpus composés de discours électoraux de trois pays, à savoir la Suisse, la France et les Etats-Unis.

3.1. Justification du choix des corpus de référence

Afin de vérifier la qualité des outils proposés, nous avons sélectionné divers ensembles de discours politiques rédigés en langues française et anglaise. Plusieurs raisons justifient ce choix. Premièrement, nous devons garantir une grande

homogénéité et nous avons donc repris uniquement des documents électoraux. Ces derniers sont rédigés afin de renforcer la motivation des partisans et de rallier d'autres électeurs, mais indiquent également les principaux choix politiques que le leader ou la formation politique entend suivre durant la prochaine législature. Le choix des mots et des expressions n'est pas le simple fruit du hasard et chaque intervenant prend un soin particulier à rédiger son intervention ou son programme. Les auteurs disposent donc d'une assez grande liberté de choix tant sur le plan du lexique, des formulations, que des thèmes retenues. En campagne électorale, chaque chef de parti peut insister ou négliger complètement telle ou telle question en recourant aux mots qu'il juge les plus pertinents. A contrario, un président ou un premier ministre doit tenir compte des diverses composantes de sa majorité ce qui entraîne des implications directes dans le choix de son lexique et de ses expressions.

Deuxièmement, comme ces discours ont été rédigés durant la même période (2007), utilisant la même langue (le français) et désirant atteindre des objectifs similaires, leur comparaison directe s'en trouve facilitée. En effet, il est connu que la comparaison entre des œuvres littéraires de genres différents mais du même auteur sont parfois plus distantes que des œuvres de même genre mais d'auteurs différents (Labbé *et al*., 2007). De plus, des études antérieures portant sur le discours politiques existent (Labbé *et al*., 2003), (Labbé *et al*., 2008) permettant une comparaison avec nos travaux. Signalons toutefois que ces dernières portent sur le discours gouvernemental et non électoral. Afin de mettre en lumière des usages récurrents ou divergents, nous avons également inclus des discours électoraux tenus lors de la dernière élection américaine (année 2008), ces derniers étant toutefois rédigés en langue anglaise.

Troisièmement, les documents disponibles seront rédigés avec un souci de garantir une qualité éditoriale. Contrairement à la *blogosphère*, nous ne pensons pas trouver de nombreuses fautes d'orthographe et d'accord ni l'usage abusif d'abréviations ou d'un vocabulaire spécifique, voire d'un langage SMS (par exemple, "sui arivé paC 11h") .

3.2. Acquisition des données

Dans le but de connaître les particularités comparatives du discours électoral suisse, nous avons constitué un corpus en téléchargeant les documents disponibles sur les sites Internet des quatre grands partis suisses. Ces textes présentent la plate-forme électorale en vue des élections fédérales d'octobre 2007. Quelques statistiques concernant ce corpus sont reprises dans le tableau 1. Nous y retrouvons la taille des quatre sous-corpus (en octets), le nombre de mots ainsi que le nombre de vocables définis comme le nombre de mots distincts.

Nous avons limité notre analyse comparative aux quatre grands partis[2] présents au Conseil fédéral (exécutif) à savoir, en partant de la droite, l'UDC (union démocratique du centre), le PRD (parti radical démocratique), le PDC (parti démocrate-chrétien) et le PS (parti socialiste). Ces formations disposent d'un nombre variable d'élus sous la Coupole fédérale. Ainsi après les élections d'octobre 2007, l'UDC dispose de 69 représentants sur 246, ce qui correspond à 28,1 % des élus. Ce parti d'une droite dure et populiste constitue la formation ayant le plus progressée lors des dernières élections (dont un gain de six sièges en 2007). Le PS reste la deuxième force politique du pays avec 52 élus (ou 21,1 %) tandis que le parti du centre, le PDC disposera de 43 élus (ou 18,7 %). Le PRD, représentant la droite modérée, doit faire face à un recul (perte de sept sièges par rapport à 2003 pour un total de 43 élus, soit 17,5 %).

	PS	**PDC**	**PRD**	**UDC**
Taille (octet)	236 821	339 047	181 381	612 134
Nb mots	35 846	50 302	26 639	90 559
Nb vocables	4 167	5 238	3 293	7 191
Documents	7	22	9	8

***Tableau 1** : Quelques statistiques sur notre corpus suisse*

Comme l'année 2007 a également connu l'élection présidentielle française, nous pouvons, dans une certaine mesure, comparer les discours électoraux tenus dans les deux pays. Recourant à la même langue, ces discours sont tenus pour des élections se déroulant dans un intervalle de six mois (avril - octobre). Ce délai assez bref nous permet d'estimer que les préoccupations des électeurs demeureront assez similaires. Afin d'atteindre cet objectif comparatif, nous avons récupéré des discours prononcés par les deux derniers candidats à l'Elysée, soit Ségolène Royal et Nicolas Sarkozy. Quelques statistiques décrivant ce corpus sont présentées dans la partie gauche du tableau 2.

	S. Royal	**N. Sarkozy**	**J. McCain**	**B. Obama**
Taille (octet)	579 323	702 627	906 161	1 644 589
Nb mots	93 479	116 212	154 665	294 553
Nb vocables	7 084	7 702	7 792	7 663
Discours	11	17	72	114

***Tableau 2** : Quelques statistiques sur le corpus français et américain*

[2] La Suisse connaît plusieurs partis de taille plus réduite dont, entre autres, le parti écologiste suisse (véritable cinquième force apparue dans les années quatre-vingt et qui renouvelle la gauche), le parti évangélique et le parti libéral. Ce dernier a fusionné avec le parti radical démocratique en octobre 2008.

Une différence importante existe tout de même entre les deux pays. Du côté français, nous sommes en présence de discours prononcés tandis que du côté suisse nous avons une forme écrite représentant une plate-forme électorale. L'absence d'un leader politique pour chaque parti suisse entraîne l'absence d'une valorisation prononcée du discours personnel. Cette variation de forme peut influencer le choix du lexique et donc avoir un impact sur nos analyses.

Finalement, nous avons également constitué un corpus en langue anglaise correspont aux discours électoraux prononcés lors de la dernière élection à la Maison Blanche (2008). Nous avons également limité notre investigation aux deux derniers candidats, à savoir les sénateurs John McCain et Barack Obama. Les statistiques décrivant ce corpus sont repris dans la partie droite du tableau 2. Dans ce dernier cas, nous avons un nombre nettement plus important de discours pour un volume textuel qui approximativement double. En revanche, le vocabulaire, mesuré en nombre de formes distinctes ou vocables, ne s'accroît pas beaucoup. On notera toutefois que la langue française dispose d'une morphologie flexionnelle nettement plus riche, rendant cette dernière comparaison directe peu judicieuse.

3.3. Prétraitement du corpus

Tout traitement de la langue naturelle nécessite un découpage en mots. Lors de la segmentation des documents, nous avons considéré comme mot toute séquence de lettres et / ou de chiffres. Cette définition laisse quelques imperfections. Ainsi, la forme “chemin de fer” sera analysée comme trois mots et les termes “ne … pas” ou “parce que” mériteraient d’être comptés sous une entrée unique. D’autre part, les formes “l’école” ou “aujourd’hui” seront vue comme composée de deux mots. En revanche, nous tiendrons compte de la langue sous-jacente et les formes “don’t” ou “I’ll” seront analysées comme deux mots dans la langue de Shakespeare.

Nous ne tiendrons pas compte de la distinction entre majuscule et minuscule. Les formes “Suisse” ou “suisse” seront considérées comme identiques. Certes, les formes “poste” et “Poste” correspondent à deux entités sémantiques distinctes dans la phrase “le poste ouvert à la Poste”. Toutefois, si un mot s’écrit exclusivement avec des majuscules, nous avons conservé cette forme en l’état car elle correspond souvent à un acronyme (“UE”, “PS”, “ONU”) ou, pour la langue anglaise, nous retrouverons les formes “US”, “NAFTA” ou “OPEC”.

```
Si la finale est ‘-ies’ mais pas ‘-eies’ ou ‘-aies’
        alors remplacez  ‘-ies’ par ‘-y’, fin;
Si la finale est ‘-es’ mais pas ‘-aes’, ‘-ees’ ou ‘-oes’
        alors remplacez  ‘-es’ par ‘-e’, fin;
Si la finale est ‘-s’ mais pas ‘-us’ ou ‘-ss’  alors éliminez  ‘-s’;
fin.
```

***Tableau 3** : Les trois règles de l’enracineur léger suggéré par Harman (1991)*

Nous n'avons pas effectué une analyse morphologique poussée afin de déterminer pour chaque mot son entrée dans le dictionnaire (lemmatisation). Dans notre cas, les formes "peux" et "pouvons" ne seront pas regroupées sous le même vocable "pouvoir". Remarquons que ce dernier peut être ambigu et que le contexte précisera s'il s'agit du nom ou du verbe. Nous avons toutefois appliqué un enracineur léger (Savoy, 2002) supprimant la marque du pluriel (le '-s' final ou la transformation de la séquence finale '-aux' en '-al'). Pour la langue anglaise nous n'avons pas retenu un enracineur agressif comme l'algorithme de Porter (1980) basé sur environ 60 règles. Nous nous sommes limités à supprimer la consonne finale '-s' indiquant souvent la forme pluriel de la langue anglaise selon l'algorithme décrit dans le tableau 3 (Harman, 1991). Une étude récente (Fautsch *et al.*, 2009) démontre que cette solution minimale propose une qualité de réponse, statistiquement comparable, à celle de Porter, en recherche d'information pour le moins.

L'application de cette procédure de normalisation nous a permis de réduire le nombre de vocables de 13 008 à 11 011 (soit une différence relative de 15,4 %) dans le cas du corpus suisse. Parfois, la forme au singulier ou au pluriel s'avère aussi fréquente l'une que l'autre (par exemple, dans le discours de l'UDC le vocable "rente" (76 occurrences) ou "rentes" (fréquence de 71)), mais plus souvent une des formes tend à dominer (par exemple, le vocable "enfants" (130 occurrences) comparée à "enfant" (9) auprès du parti PDC). En règle générale, nous pensons que les calculs effectués divergent quelque peu par rapport à une lemmatisation complète, mais les conclusions que nous en tirons devraient demeurer identiques, très similaires pour le moins.

Finalement, nous garderons à l'esprit que le choix du lexique sera sujet à des variations dues aux circonstances (le contexte, l'auditoire, intervention spontanée ou discours lu) ainsi qu'au type de communication choisi (programme général ou discours traitant une question particulière ou technique). Dans le contexte présent, ces diverses variations sont relativement neutralisées dans nos corpus. En effet, les textes proviennent de la même période, sont rédigés dans la même perspective et couvrent des objectifs très similaires. En revanche, si l'on passe d'un pays à l'autre ou d'une langue à une autre, les comparaisons devront tenir compte des différences entre les électorats et les cultures politiques.

4. Méthodes d'extraction et de représentation

L'indexation automatique vise à définir l'importance de chaque terme et des les pondérer en tenant compte essentiellement de leur fréquence lexicale (*tf*), de la fréquence documentaire d'un terme (*df*, ou plus précisément de l'*idf* = $\log(n/df)$) et de la longueur du document (Boughanem *et al.*, 2008). Dans notre contexte, le nombre de documents ou de sites distincts demeure faible (n = 4 pour le corpus suisse et deux pour les corpus français ou américain) réduisant l'intérêt pour une mesure *idf*. En effet de nombreux mots apparaissent dans tous les sites et leur valeur

idf sera donc nulle. La suite de cette section examine les possibilités d'extraire les éléments comparatifs d'un site (ou d'un ensemble de documents) en fonction d'autres corpus de référence.

4.1. Richesse lexicale et vocabulaire

L'ampleur du vocabulaire utilisé par les différents orateurs pourrait constituer une première mesure. Dans cette optique, nous pouvons estimer qu'un lexique étendu correspond à un candidat ou à un parti ayant de grandes ambitions, désirant couvrir tous les domaines (Labbé *et al*., 2003). A contrario, en présence d'un volume lexical plus restreint, nous pourrions avancer l'hypothèse que le candidat ou le parti a opté pour la sobriété, pour un parler simple et direct, une communication qui se veut plus proche du peuple et dans un souci d'éviter toute formulation trop sophistiquée. Cependant, cette analyse doit s'effectuer sur un ensemble de documents possédant la même taille ou, à défaut, de longueur très similaire. En effet, un corpus possédant un volume plus important proposera également un vocabulaire plus ample et sera donc ainsi favorisé (Baayen, 2001). Comme l'indiquent les tableaux 1 et 2, les quatre grands partis suisses ou les candidats aux élections présidentielles présentent des volumes assez différents. Afin de normaliser les divers corpus, nous les avons réduit à la taille du plus faible, soit celui du PRD pour le corpus suisse (voir tableau 4), et celui de S. Royal et de J. McCain pour, respectivement, les corpus français et américain (voir tableau 5).

	PS	**PDC**	**PRD**	**UDC**
Nb de mots	26 639	26 639	26 639	26 639
Nb vocables	3 412	3 682	3 293	3 899
Hapax	1 676	1 811	1 511	1 964
Hapax en %	49,1 %	49,2 %	45,9 %	45,3 %

***Tableau 4** : Richesse lexicale du corpus suisse en nombre de vocables (forme distincte) et nombre de vocables n'apparaissant qu'une seule fois (hapax)*

Sur la base du tableau 4, nous pouvons comparer le nombre de vocables utilisés par les quatre grands partis suisses. Dans ce cas, le vocabulaire le plus étendu se rencontre auprès de l'UDC avec 3 899 formes, suivi par le PDC (3 682 vocables), puis le PS (3 412) et, finalement, le PRD (3 293). Le parler simple et direct serait l'apanage du PRD tandis que les grandes ambitions et la couverture la plus large seraient plutôt du côté de l'UDC. D'un autre côté, si la rareté des mots est un indice de la richesse lexicale avec des expressions savantes n'apparaissant qu'une seule fois, l'UDC remporte de nouveau le premier rang avec 1 964 vocables n'apparaissant qu'une seule fois (*hapax legomena*), contre 1 811 pour le PDC, 1 676 pour le PS et 1 511 pour le PRD. Ces informations nous fournissent un indice lexical global mais pas une indication précise de la sémantique sous-jacente.

	S. Royal	N. Sarkozy	J. McCain	B. Obama
Nb de mots	93 469	93 469	154 365	154 365
Nb vocables	7 084	7 074	7 792	6 679
Hapax	3 124	3 143	2 958	2 417
Hapax en %	44,1 %	44,4 %	38,0 %	36,2 %

***Tableau 5** : Richesse lexicale des corpus français et américain en nombre de vocables (forme distincte) et nombre d'hapax*

Comme l'illustre le tableau 5, les deux candidats à l'Elysée ne se distinguent pas sur ce critère de richesse lexicale, chacun ayant un nombre de vocables ou d'*hapax* très similaire. Lors de l'élection à la Maison Blanche, le candidat républicain possède un lexique plus vaste que B. Obama (7 792 vocables contre 6 679). De plus, l'usage de mots plus rares confirme cette tendance, le démocrate semblant opter pour une plus grande répétition des mots et des expressions favorisant un parler simple et direct.

4.2. Fréquence d'occurrence

Afin de déterminer les mots correspondant au contenu sémantique d'un document, on peut recourir à la fréquence d'apparition (*tf*). Parmi les formes très fréquentes, nous pouvons alors observer les mêmes termes et ceci quelle que soit la formation politique. Un regard plus attentif révèle que ces vocables abondants correspondent à des mots outils ("de", "la", "et", "en", "une", "dans", "pour", etc.) ou, pour la langue anglaise ("the", "and", "to", "that", "of", "a", etc.). Après élimination de 64 termes français (ou de 258 mots anglais) peu porteurs de sens, nous voyons mieux émerger les thèmes récurrents et communs à l'ensemble des discours et par différence, ceux propres à chaque parti ou candidat.

Le tableau 6 indique pour chaque parti suisse les dix vocables les plus fréquents. A côté de chaque forme, nous avons noté sa fréquence d'occurrence (*tf*). Nous pouvons constater que certains mots apparaissent fréquemment dans les quatre discours comme "politique", "suisse", "être" et "ne … pas" ou, dans trois des quatre, comme "doit" ou "nous". Certaines formes apportent peu d'information ("être", "ne … pas", "doit", "nous") tandis que d'autres laissent clairement voir l'origine commune du corpus ("politique", "suisse"). Si l'on analyse le vocable "suisse", on constate que son rang diverge entre les partis. Pour l'UDC, ce terme s'avère le plus usité avec une fréquence d'occurrence (864) nettement supérieure au deuxième vocable le plus fréquent (vocable "pas", fréquence de 456). Pour les deux partis du centre-droit, ce terme "suisse" apparaît au deuxième rang, tandis que ce vocable semble moins utilisé au PS (septième rang).

PS		PDC		PRD		UDC	
tf	vocable	*tf*	vocable	*tf*	vocable	*tf*	vocable
237	nous	643	nous	178	être	864	suisse
198	politique	347	suisse	176	suisse	456	pas
192	doit	261	pas	166	doit	445	politique
190	pas	245	être	143	politique	384	ne
178	ne	230	notre	138	nous	323	être
150	être	222	ne	108	sécurité	321	état
133	suisse	177	politique	108	ne	320	AI
132	culture	174	PDC	91	pas	295	droit
106	culturelle	156	doit	90	doivent	286	UDC
104	sociale	144	formation	88	armée	248	étranger

***Tableau 6** : Les dix vocables les plus fréquents (et leur fréquence absolue) dans les discours des quatre partis suisses*

Parmi les dix vocables les plus fréquents, nous pouvons voir se dessiner des tendances propres à chaque formation politique. Ainsi, le PS semble se distinguer par l'usage fréquent du vocable "culture" ainsi que la forme reliée "culturelle" tandis que le PDC recourt volontiers à "formation". A droite, le PRD se positionne sur les thèmes de la "sécurité" et de l'"armée" tandis que l'UDC insiste sur les vocables "droit", "étranger" et les problèmes de l'assurance-invalidité (AI).

S. Royal		N. Sarkozy		J. McCain		B. Obama	
tf	vocable	*tf*	vocable	*tf*	vocable	*tf*	vocable
1380	je	1862	je	2345	I	6203	we
860	nous	1663	pas	2160	we	4216	I
744	vous	1267	ne	1602	our	3575	not
633	pas	851	france	1540	will	3276	our
539	france	708	nous	1323	not	3164	will
527	ne	575	veux	821	my	2389	you
401	avec	503	si	775	you	1566	american
348	leur	498	leur	775	american	1313	can
335	tous	480	parce	640	he	1107	america
296	aussi	448	être	540	country	1081	year

***Tableau 7** : Les dix vocables les plus fréquents (et leur fréquence absolue) dans les discours des candidats à l'Elysée et à la Maison Blanche*

Pour les candidats à la présidence française ou américaine (voir tableau 7), nous pouvons retrouver les mêmes tendances entre eux et similaires à celles décelées du

côté helvétique (voir tableau 6). Dans tous les cas, les formes négatives ("ne … pas" ou "not") s'avèrent très fréquentes, ainsi que le pronom "nous" ou "we" sans que ces éléments ne reflètent, a priori, une information très pertinente. Mais d'autres vocables fournissent des renseignements un peu plus intéressants.

D'abord on notera que l'origine géographique des corpus s'identifie clairement (vocable "france" ou "american" et "america"), bien que la provenance politique n'apparaissent que plus loin dans la liste (par exemple, par le mot "politique", 17e chez S. Royal, 13e chez N. Sarkozy ou "république", 79e chez S. Royal, 25e chez N. Sarkozy). Dans le corpus américain, l'emphase sur la politique apparaît avec les vocables "senator" (16e chez J. McCain ou 39e chez B. Obama), voire le mot "government" chez J. McCain (13e position) et "Washigton" chez B. Obama (34e rang). Chez B. Obama, le mot "government" est nettement moins usité, n'apparaissant qu'en 82e position.

Avec un peu d'étonnement, le but de l'élection reste peu visible du côté français. Ainsi le vocable "président" apparaît seulement en 94e position chez N. Sarkozy tandis que c'est la forme "(pacte) présidentiel" qui apparaît chez S. Royal (en position 111). Pour la campagne américaine, le vocable "president" s'avère plus usité, occupant le 23e rang chez J. McCain et le 24e chez B. Obama.

Ensuite, les petits mots font parfois toute la différence et dans ce cas nous rencontrons un emploi marqué du "je" (ainsi que du vocable relié "j", "me" ou "moi") dans les discours électoraux français ou américain ("I", "me" ou "my") par rapport à ceux de la Suisse. Ce vocable "je" indique bien l'importance attachée à une personne, au chef du parti ou au futur président. Plus étonnant, la fréquence d'occurrence du pronom "je" s'avère statistiquement plus élevée pendant le deuxième tour de la campagne présidentielle française que lors du premier tour (Labbé *et al*., 2008b). Le passage au second tour s'accompagne bien d'un changement au niveau lexical et cela se vérifie pour les deux candidats. Une campagne ne forme pas un continuum lexical stable, mais des ruptures peuvent apparaître. En fin de course, il faut serrer les rangs autour du "moi", du chef qui insistera sur le "je veux" ("(as) President, I will"). Ce vocable peut également s'expliquer, en partie, par la forme orale du discours électoral français ou américain, restant complètement absent du côté helvétique.

Un politologue anglo-saxon notera que l'usage du vocable "I" s'avère souvent attaché au premier ministre durant sa campagne électorale, tandis que le mot "we" s'utilisera plus par le chef de l'opposition. Selon cette perspective, J. McCain serait plus le chef du gouvernement défendant son bilan et B. Obama le chef de l'opposition proposant des changements.

Enfin, les discours électoraux français et américain connaissent un souci explicatif indéniable avec le vocable "parce (que)" ou "because" (13e chez B. Obama, mais seulement 47e chez J. McCain). Les formes verbales fréquentes correspondent au verbe "vouloir" ("veux", 6e chez N. Sarkozy, 11e chez S. Royal). Du côté américain, on retrouve ces formes qui indiquent la volonté et le pouvoir de

les réaliser (“will”, “can” (15e chez J. McCain), voire “make” (20e chez Obama, 21e chez J. McCain)).

Comme première mesure, la fréquence d'occurrence permet de faire ressortir les vocables décrivant le contenu sémantique d'un document. Nous pourrions tenir compte de cette information pour définir l'importance de chaque terme dans une représentation (par exemple, par des variations de fonte, taille ou de couleur dans une interface de type “nuage de termes”). On prendra soin toutefois de normaliser cette fréquence lexicale en fonction de la longueur du document sous-jacent (par exemple, *tf*/max *tf*).

Comme alternative, on peut retenir la fréquence lexicale uniquement comme clé de tri. L'importance de chaque terme se mesurerait alors uniquement en fonction de leur rang. Ce second choix ne s'avère pas toujours très satisfaisant car une différence de rang unitaire peut cacher des situations très différentes[3]. Par exemple, la différence de rangs entre les vocables “pas” et “suisse” pour le parti UDC (voir tableau 6) s'élève à 1, tandis que la différence de fréquence se monte à 408 (864 - 456) ou de manière relative à 0,472 (= 864/864 - 456/864).

4.3. Représentation comparative

Afin de déterminer le vocabulaire spécifique à une personne ou à un parti, nous devons le comparer à une norme, par exemple à un corpus composé d'un ensemble de personnes ou de partis dans des contextes similaires. Comme les tableaux 6 et 7 l'indiquent, plusieurs vocables apparaissent fréquemment parmi plusieurs orateurs et ne permettent donc pas de discriminer clairement les contenus respectifs. Avec un corpus de référence, nous pouvons observer quelle forme apparaît de manière significativement plus fréquente dans l’un ou l’autre des discours ou, inversement, ceux dont l’occurrence s’avère significativement moins forte. Pour atteindre cet objectif, nous nous sommes inspirés de la méthode proposée par Muller (1992).

Regroupons tous les documents que l'on désire comparer pour former un corpus général **C**. Si nous désirons définir une représentation comparative par rapport à cette norme **C**, nous pouvons extraire les documents correspondant à un sous-ensemble noté **S**. Le reste du corpus sera noté **C-** (avec $\mathbf{C} = \mathbf{S} \cup \mathbf{C}$-).

Fixons un vocable particulier noté ω. Nous pouvons alors compter sa fréquence d'apparition dans le sous-ensemble **S** (valeur notée *a* dans le tableau 8) et sa fréquence dans le reste du corpus **C-** (valeur notée *b*). Evidemment, la fréquence d'occurrence dans l'ensemble **C** sera de $a + b$. De manière similaire, nous pouvons compter la fréquence lexicale de tous les autres vocables dans le sous-ensemble **S** (valeur notée *c*) et le reste du corpus **C-** (fréquence notée *d*). Le corpus **C** comprendra donc *n* mots avec $n = a + b + c + d$.

[3] La distribution de la fréquence d'occurrence suit une loi de puissance (Baayen, 2001).

	S	**C-**	
ω	*a*	*b*	*a* + *b*
différent de ω	*c*	*d*	*c* + *d*
	a + *c*	*b* + *d*	*n* = *a* + *b* + *c* + *d*

***Tableau 8** : Exemple d'une table de contingence pour le vocable ω*

Ensuite, nous faisons l'hypothèse que le mot ω suit une distribution binomiale avec comme paramètres p et n'. Le paramètre p indiquant la probabilité d'occurrence du terme ω . Cette dernière peut être estimée par $p = (a+b) / n$ tandis que $n' = a+c$ correspond à la taille (en nombre de mots) du sous-ensemble **S**. Cette estimation d'une probabilité suit le principe du maximum de vraisemblance qui conduit à surestimer les probabilités des vocables présents au détriment des vocables absents. Dans ce dernier cas, la valeur $a+b$ serait égale à 0, donnant une probabilité nulle d'occurrence. Or, il est reconnu que la distribution des mots suit une distribution de type LNRE (*Large Number of Rare Events* (Baayen, 2001)). Comme correction, un lissage simple (Manning *et al*., 2000) consiste à ajouter une unité au numérateur de notre estimation et, en complément, d'ajouter au dénominateur la taille du vocabulaire retenu. Cette formulation se généralise (loi de Lidstone) en lissant toute probabilité par la formule $p = (a+b+\lambda) / (n'+(\lambda \cdot |V|))$, avec λ un paramètre de lissage (fixé à 0,3) et |V| indiquant la taille du vocabulaire (*par exemple*, 10 950 pour le corpus suisse, 10 678 pour les discours français ou 10 410 pour le corpus américain). Comme alternative, l'approche suggérée par Good-Turing (Sampson, 2001) redonne, habituellement, des estimations plus fiables pour les faibles fréquences lexicales (Gale *et al*., 1994). Or, notre centre d'intérêt ne concerne pas *a priori* les fréquences peu élevées (entre un et trois) et l'implémentation du lissage Good-Turing demeure plus complexe à mettre en œuvre.

Notre modèle expliquant les fréquences lexicales suivant une distribution binomiale, nous pouvons calculer le nombre moyen attendu d'apparition du terme ω dans la partie **S** qui s'élève à $n' \cdot p$, avec comme écart type la valeur $n' \cdot p \cdot (1-p)$. Or le tableau 8 indique le nombre réellement observé dans la partie **S** qui est noté a.

La différence entre le nombre réellement observé a et la moyenne théorique $(n' \cdot p)$ permet de déterminer les suremplois (différence positive) et les vocables sous-représentés (différence négative). Mais nous devons encore tenir compte de la variabilité sous-jacente mesurée par l'écart type. Nous devons donc calculer une moyenne normalisée ou un score normalisé (noté score Z) pour chaque terme ω selon l'équation 1.

$$\text{score Z}(\omega) = \left[\frac{a - n' \bullet p}{\sqrt{n' \bullet p \bullet (1-p)}} \right] \quad (1)$$

Dans cette formule, nous tenons compte du nombre observé d'occurrences (valeur a) auquel on soustrait sa moyenne théorique (nombre prédit par la

distribution binomiale). Cette différence doit encore être divisée par un estimateur de l'écart-type pour retourner une valeur standardisée.

Enfin, comme règle de décision, nous pouvons décider qu'un score Z supérieur à 3 indiquera un suremploi significatif[4] du vocable, tandis que des valeurs négatives et inférieures à -3 signalent des sous-emplois. En faisant varier cette limite, nous pouvons extraire un nombre variable de vocables comme descripteurs comparatifs d'un document ou d'un sous-corpus. Comme approche alternative, nous pouvons trier les vocables selon leur score Z et extraire les *k* formes ayant la plus forte valeur, sous condition que cette dernière soit supérieure à 2. Cette dernière option sera utilisée dans la prochaine section pour l'analyse des discours politiques.

5. Applications en analyse des discours électoraux

Sur la base des discours électoraux suisses, nous avons pu déterminer les termes caractéristiques des quatre formations (section 5.1). Comme l'année 2007 a également connu l'élection présidentielle française, et l'année 2008 la campagne électorale américaine, nous avons analysé les représentations possibles des deux derniers candidats pour ces deux élections (section 5.2). Enfin, dans la section 5.3 nous avons décidé de comparer les discours électoraux tenus dans les deux pays francophones et ainsi de détecter le vocabulaire spécifique aux deux élections.

5.1. Application aux discours des partis suisses

Sur la base de notre méthodologie décrite en section 4.3, nous avons déterminé les vocables sur-employés pour chaque formation politique (tableau 9), ainsi que ceux qui sont sous-représentés (tableau 10). Pour établir ces listes ordonnées, nous avons calculé le score Z normalisé de chaque terme, valeur indiquée à côté de chaque entrée dans les tableaux 9 et 10. Le corpus de référence est constitué des plates-formes électorales de l'ensemble des quatre partis suisses.

Les vocables apparaissant dans le tableau 9 forment une représentation comparative des thèmes privilégiés par chaque formation. Elle s'avère plus parlante et pertinente qu'une représentation s'appuyant uniquement sur la fréquence d'occurrence (voir tableau 6). Pour l'extrême droite (UDC), les thèmes récurrents sont les assurances sociales ("rentes", "AI"), la politique de naturalisation, la neutralité de la Suisse et la défense de son identité face à l'étranger, l'affectation des ressources financières ("milliard", "franc" (11e)), mais également le souci de se distinguer de la "gauche". Pour la droite modérée (PRD), les sujets touchant la sécurité ("armée", "défense", "sécurité", "mission", "militaire") forment une

[4] En admettant que la valeur *Z* suit une distribution normale, les valeurs excédant les limites de 3 et -3 représentent 0,3 % des cas. En descendant cette limite à 2 et -2, on trouverait théoriquement 4,6 % des observations.

thématique centrale ainsi que les questions d'imposition fiscale ("*easy*" et "*tax*" dans l'expression "*easy swiss tax*"). Le parti du centre (PDC) axe son discours sur la famille ("enfant") mais de manière un peu surprenante sur l'énergie et l'environnement ("énergie", "énergétique") d'une part et, d'autre part, sur la technologie ("internet", "technologique" (11e) "électronique" (12e)). Le parti socialiste (PS) semble se caractériser par sa politique culturelle ("culture", "artiste", "art" ou "pro" dans la dénomination "pro helvetia"), mais également par une préoccupation écologique (taxe sur le "CO2") à côté d'un thème plus traditionnel ("autogestion")[5].

PS		**PDC**		**PRD**		**UDC**	
Z	vocable	Z	vocable	Z	vocable	Z	vocable
15,2	état	21,8	nous	18,9	PRD	14,6	AI
14,0	II	18,9	PDC	16,0	radical	13,2	UDC
13,0	culture	11,8	demandons	12,2	mission	11,3	neutralité
11,9	culturelle	10,4	énergie	12,0	armée	10,0	gauche
11,7	artiste	10,1	internet	11,7	défense	9,6	naturalisation
10,3	encouragement	9,1	enfant	11,3	sécurité	9,0	rente
10,1	art	9,1	notre	9,6	militaire	8,8	état
10,0	autogestion	8,9	énergétique	9,6	*easy*	8,7	nationalité
10,0	CO2	8,2	thème	9,5	imposition	8,0	milliard
9,5	pro	8,1	jeune	9,2	*tax*	7,4	suisse

***Tableau 9** : Les dix vocables les plus surreprésentés dans les sites des partis suisses*

Ces vocables ne sont pas forcément très fréquents. Ainsi, le terme "easy" apparaît 16 fois et uniquement dans les discours du PRD, tandis que l'on compte 26 occurrences du terme "autogestion" utilisé uniquement dans le programme du PS. On se gardera d'en tirer la conclusion que les termes surreprésentés apparaissent seulement auprès d'un auteur. Ainsi, on compte 19 occurrences du vocable "tax" mais 17 fois dans le discours du PRD ce qui en fait un terme sur-employé pour cette formation.

De manière duale, les vocables peu usités dans chaque formation politique permettent de compléter ces conclusions (voir tableau 10). Ainsi, on constate que les sigles des autres partis ne sont que très peu fréquents dans le discours d'un parti donné. On ne compare pas son programme avec les autres et on se garde bien de mentionner les autres à l'exception de l'UDC avec ses vocables "PS" et "gauche". Les termes "neutralité" ou "AI" sont visiblement des termes propres à l'UDC. Etonnamment, le terme "neutralité" est sous-employé par le parti PRD dont les

[5] Le vocable « II » surreprésenté dans le discours du PS correspond au pronom « il ». Pour une raison inconnue, la forme « II » a été substituée au pronom « il » dans les documents disponibles sur Internet et décrivant la plate-forme de ce parti.

thèmes majeurs concernaient la sécurité et l'armée. Enfin, le PS n'utilise que fort peu le terme "suisse" ou "état" mais également "UE", lui qui est le seul parti à souhaiter l'ouverture de négociations en vue de l'adhésion à l'UE.

PS		**PDC**		**PRD**		**UDC**	
Z	vocable	Z	vocable	Z	vocable	Z	vocable
-8,2	suisse	-8,7	AI	-6,3	UDC	-17,2	nous
-8,1	état	-7,2	neutralité	-5,9	AI	-8,1	PDC
-7,6	AI	-7,2	UDC	-5,5	gauche	-7,5	notre
-6,9	UDC	-7,2	culture	-5,2	culture	-6,4	voulons
-5,9	gauche	-6,9	culturelle	-5,1	franc	-5,9	économique
-5,6	enfant	-6,5	armée	-4,9	PDC	-5,7	demandons
-5,3	jeune	-6,0	naturalisation	-4,9	PS	-5,5	formation
-5,3	neutralité	-6,0	rente	-4,6	ont	-5,4	état
-5,2	école	-5,4	PS	-4,5	neutralité	-5,3	II
-5,1	UE	-5,3	nationalité	-4,5	année	-5,1	cadre

***Tableau 10** : Les dix vocables les plus sous-employés par les partis suisses*

5.2. Elections présidentielles française et américaine

Afin de mieux connaître le vocabulaire spécifique des élections de part et d'autre de l'Atlantique, nous avons regroupé les vocables sur-employés pour chaque candidat dans le tableau 11. Le corpus de référence est constitué de l'ensemble des discours électoraux pour chaque pays, évidement pris séparément.

Entre les candidats de chaque élection, nous voyons mieux se dessiner les différences. Ainsi, du côté socialiste on retrouve une partie de leurs thèmes sous les vocables "salarié", "social", "énergie", "développement" et "durable". On remarquera également la personnalité du chef et son implication dans le lexique de la campagne avec les termes "pacte", "présidentiel" ou "femme". Mais également le souci d'établir un lien avec le public (vocable "vous") et l'accent placé sur les préoccupations présentes ("aujourd'(hui)"). Plus loin dans la liste des vocables sur-employés, on retrouve les termes "logement" (18e rang), "environnement" (20e rang) et "entreprise" (23e rang).

Pour N. Sarkozy, on notera l'usage de la négation ("ne … pas") et la nécessité de l'explication ("parce (que)"), souci que l'on retrouvera chez B. Obama ("because"). L'identité nationale ("français") (la forme "identité" se trouve en 17e position des sur-emploi), la culture et la volonté de charger ("veux") complètent les sur-emplois. Comme opposition de style, on notera l'abondance de la forme "homme" (qui se retrouve seulement en 167e position dans le discours électoral de S. Royal). Comme éléments complémentaires, on retrouve, chez le leader de l'UMP, les termes "parler"

(16e rang), “rêve” (20e rang) et “voyou” (22e rang, mot absent du lexique de S. Royal). Chez le futur président, on constate l'usage plus abondant de verbes tandis que S. Royal usait plus volonté de noms.

S. Royal		**N. Sarkozy**		**J. McCain**		**B. Obama**	
Z	vocable	Z	vocable	Z	vocable	Z	vocable
10,55	vous	12,52	pas	20,80	Obama	11,46	we
9,85	pacte	9,99	ne	12,47	government	10,84	McCain
8,03	jeune	6,90	culture	8,71	my	7,91	you
7,65	femme	6,59	si	8,47	elected	7,90	because
7,33	état	6,18	françai	7,95	dollar	7,65	street
7,24	là	6,10	parce	7,72	greater	7,19	John
6,65	salarié	5,61	unique	7,61	spending	6,54	Bush
6,58	nous	5,40	pensée	7,55	nuclear	6,39	college
6,48	sera	5,40	rien	7,36	low	6,26	afford
6,42	présidentiel	5,32	morale	7,17	federal	6,00	want
6,13	social	5,31	outremer	6,97	public	5,91	can
6,04	aujourd	5,28	était	6,78	court	5,85	change
6,02	énergie	5,28	homme	6,59	power	5,70	class
6,01	durable	5,17	veux	6,46	case	5,39	century
5,98	développement	5,15	être	6,38	Canada	5,29	time

***Tableau 11** : Les quinze vocables les plus surreprésentés lors des dernières élections présidentielles en France et aux Etats-Unis*

Avec l'élection américaine, d'autres usages apparaissent. Ainsi, les sur-emplois des vocables “Obama”, ou “John” “McCain” démontrent que chaque candidat fait référence directement à l'autre et, inversement, n'utilise pas ou peu son nom dans ses propres allocutions. Le style possède aussi son influence sur le lexique. Ainsi, le candidat républicain préfère parler de “government” ou de “(if I'm) elected” tandis que le démocrate fera référence à l'administration “Bush” avec la nécessité de changer les choses (et les slogans “(yes we) can” “(the) change (we belive in)”). Au niveau des thématiques récurrentes, on voit également apparaître les différences. Du côté de J. McCain, des propositions pour un renforcement de l'énergie atomique (“nuclear”, “power”), des tribunaux fédéraux (“federal”, “court”) et le souci de maîtriser les dépenses du gouvernement fédéral (“dollar”, “spending”). Le futur président B. Obama va insister sur les problèmes de l'économie réelle (“(main) street”) et de la finance (“(wall) street”). De plus on voit se dessiner sa volonté de renforcer l'éducation (“college”) et les autres besoins et demandes (comme la santé) que le peuple et en particulier la “(middle) class” ne peut plus s'offrir (“afford”). Finalement quelques formes peuvent révéler des aspects secondaires des candidats

comme, par exemple, le fait que le sénateur J. McCain a prononcé plusieurs discours au Canada.

5.3. Spécificité du vocabulaire électoral suisse et français

Dans notre représentation et analyse comparative, nous pouvons regrouper l'ensemble des documents rédigés en langue française. Sur cette base, nous pouvons mieux distinguer le vocabulaire spécifique à une élection comme l'illustre le tableau 12. En premier lieu, on y retrouve les dénominations propres à chaque pays ("suisse", "france" et "francai(s)") de même que celles rattachées à leurs institutions respectives, à savoir "fédéral(e)", "confédération", "canton", "conseil", "UDC", "PDC" d'une part et "république" ou "nation" (en 27e position) d'autre part.

Deuxièmement, comme souligné précédemment, le discours français connaît une surabondance du pronom "je" (ou "j", "me", "moi" en 15e rang) et la volonté d'impliquer l'auditeur avec le pronom "vous". Du côté helvétique, pas de présence significative de ces pronoms personnels.

Suisse			France		
Z	tf	vocable	Z	tf	vocable
28,94	1520	suisse	40,08	3242	je
15,21	424	fédéral	26,53	1390	france
13,38	326	AI	23,21	1156	vous
12,92	307	confédération	21,00	867	veux
12,85	301	UDC	17,42	607	françai
12,81	367	étranger	16,93	680	parce
11,91	270	franc	16,74	604	j
11,87	458	doivent	16,40	2296	pas
11,06	756	doit	15,34	488	ai
10,99	226	cantons	13,19	351	république
10,73	213	neutralité	12,63	1794	ne
10,73	210	UE	12,33	322	suis
10,66	285	conseil	12,10	406	dire
10,23	194	fédérale	11,89	290	me
10,12	187	PDC	11,71	485	ceux

***Tableau 12** : Les quinze vocables les plus surreprésentés dans la dernière élection fédérale (octobre 2007) et présidentielle française (avril 2007)*

Troisièmement, et de manière plus profonde, on retrouve, du côté français, les formes verbales "veux", "ai", "suis", "dire", "dis" (en 20e rang), "crois" (en 22e rang) ou la conjonction "parce que" phénomène qui s'explique, en partie, par le fait

que le discours était oral. Du côté helvétique, les formes verbales abondantes sont "doit" ou "doivent" soulignant les obligations ou attentes ("l'Etat doit"), le besoin de quantifier ("franc" (7e rang), "milliard" (19e rang)), ainsi que les vocables "assurance" (18e rang), "naturalisation" (22e rang), "économie" (25e rang) ou "culturelle" (28e rang).

Finalement, il est également intéressant de noter que l'acronyme "UE" s'avère sur-employé dans le discours politique suisse. Les deux candidats à l'Elysée n'ont pas retenu cette forme et ont préféré parler d'"Europe"[6], forme apparaissant au 90e rang selon le score Z chez S. Royal. Chez N. Sarkozy, ce vocable occupe le 13 573e rang, représentant dans ce cas un sous-emploi significatif dans le discours du futur président de la République française.

6. Conclusion

Afin de déterminer les termes comparativement les plus représentatifs, nous avons étudié la possibilité de recourir à la fréquence lexicale ou au rang. Cette information permet certes de se faire une idée du contenu d'un document ou d'un corpus. Toutefois, cette approche ne dispose pas d'une règle de décision claire et ne permet pas de distinguer entre les vocables fréquents et ceux qui caractérisent comparativement un sous-ensemble donné.

Notre méthode propose de recourir à un score normalisé (ou score Z) sous l'hypothèse que la distribution des fréquences d'occurrence suit une loi binomiale. En appliquant cette méthode pour l'analyse comparative des discours électoraux en Suisse (élection fédérale d'octobre 2007), nous pouvons mettre en lumière les thèmes porteurs et les caractéristiques des quatre grandes formations suisses. En comparaison avec la liste des vocables les plus fréquents (tableau 6), les sur-emplois, repris dans le tableau 9, laissent plus clairement apparaître les thèmes propres à chaque formation politique. De manière similaire, pour les discours électoraux français et américains, la fréquence absolue d'occurrence (tableau 7) indique quelques lignes directrices. En revanche les scores normalisés (tableau 11) font mieux ressortir les points de divergence de style et de lexique.

Par rapport à la campagne présidentielle française (avril - mai 2007), les discours politiques suisses ne recourent que très peu à la forme "je" ou aux vocables "veux", "dire", "crois" ou "parce que" dénotant l'importance du chef unique et un souci explicatif indéniable du côté français. Cette personnalisation se retrouve également aux Etats-Unis. Dans ce dernier cas, on désignera clairement son adversaire en utilisant de façon significative son nom dans les allocutions ("McCain", "Obama"). Dans le discours helvétique, un seul parti utilise des comparaisons directes avec un

[6] Le vocable « europe » apparaît en 219e position des formes les plus fréquentes des discours politiques suisses.

autre. Parmi les formes verbales significatives du discours suisse, on retrouve “doit” ou “doivent” soulignant plutôt les obligations ou attentes (“l’Etat doit”).

Enfin, la méthode proposée permet également de traiter des bigrammes (“adhésion UE”, “taxe CO2”, “pacte présidentiel”, “Wall Street”, “Main Street” ou “oil company”) ou trigrammes (“camp rouge-vert”, “nous autres radicaux” ou “je m’engage”) permettant peut-être de mieux refléter la sémantique sous-jacente. De plus, si nous avons retenu les formes de surface, nous pourrions sans difficulté appliquer la même approche sur des lemmes. Dans ce cas, des formes différentes, mais reliées au même lemme seraient réunies sous la même entrée. La solution proposée demeure simple à appliquer et ne requiert pas de corpus d'entraînement (pour construire, par exemple, un modèle de langue).

Remerciements

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

6. Bibliographie

Baayen H.R. *Word Frequency Distributions*, Dordrecht, Kluwer Academic Publishers, 2001.

Baayen H.R., van Halteren H., Neijt A. & Tweedie F. « An experiment in authorship attribution », Actes *JADT 2002*, St Malo, 13-15 mars 2002, p. 69-75.

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

Berger A.L. & Mittal V.O. « OCELOT: A system for summarizing web pages », *Proceedings ACM-SIGIR-2000*, Athens, 24-28 July 2000, New York, The ACM Press, 2000, p. 144-151.

Boughanem M. & Savoy J. *Recherche d'information. Etat des lieux et perspectives*, Paris, Hermès, 2008.

Carpenter R.H. & Seltzer R.V. « On Nixon's Kennedy Style », *Speaker and Gavel*, vol. 7, nº 41, 1970.

Fautsch C. & Savoy J. « Stratégies de recherche dans la blogosphère », *Document Numérique*, vol. 11, nº 1-2, 2008, p. 109-132.

Fautsch C. & Savoy J. « Algorithmic stemmers or morphological analysis: An evaluation », *Journal of the American Society for Information Sciences & Technology*, vol. 60, nº 8, 2009, p. 1616-1624.

Fuller M., Tsagkias M., Newman E., Besser J., Larson M., Jones G.J.F. & de Rijke M. « Using term clouds to represent segment-level semantic content of podcasts », *Proceedings 2nd SIGIR Workshop on Searching Conversational Speech*, Singapore, 24 July 2008, New York, The ACM Press, 2008.

Gale W.A. & Church K.W. « What is wrong with adding one? », In N. Oostdijk, P. de Hann (Eds), *Corpus-Based Research into Language*. Harcourt Brace, 1994.

Goldstein J., Kantrowitz M., Mittal V. & Carbonell J. « Summarizing text documents », *Proceedings ACM-SIGIR-99*, Berkeley, 15-19 August 1999, New York, The ACM Press, 1999, p. 121-128.

Harman D. « How effective is suffixing? », *Journal of the American Society for Information Science*, vol. 42, nº 1, 1991, p. 7-15.

Herman V. « What governments say and what governments do: An analysis of post-war Queen's speeches », *Parliamentary Affairs*, vol. 28, nº 1, 1974, p. 22-31.

Konchady M. *Text Mining Application Programming*, Boston, Ch. River, 2006.

Kupiec J., Pedersen, J. & Chen, F. « A trainable document summarizer », *Proceedings ACM-SIGIR-95*, Seattle, 9-13 July 1995, New York, The ACM Press, 1995, p. 68-73.

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

Labbé C. & Labbé D. « Baudelaire, Rimbaud et Verlaine », *Actes Aspects linguistiques du texte poétique*, Brest, 16-17 novembre 2007.

Labbé D. « Experiments on authorship attribution by intertextual distance in English », *Journal of Quantitative Linguistics*, vol. 14, nº 1, 2007, p. 33-80.

Labbé D. & Monière D. *Les mots qui nous gouvernent. Le discours des premiers ministres québécois : 1960-2005*, Montréal, Monière-Wollank, 2008.

Labbé D. & Monière D. « Je est-il un autre ? », Actes *JADT 2008*, Lyon, 12-14 mars 2008, p. 647-656.

Mani I. & Maybury M.T. *Advances in Automatic Text Summarization*, Cambridge (MA), The MIT Press, 1999.

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

Monière D. & Labbé D. « L'influence des plumes de l'ombre sur les discours des politiciens », Actes *JADT 2006*, Besançon, 19-21 avril 2006, p. 687-696.

Muller C. *Principes et méthodes de statistique lexicale*, Paris, Honoré Champion, 1992.

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

Orasan C., Pekar V. & Hasler L. « A comparison of summarization methods based on term specificity estimation », *Proceedings of Language and Resources and Evaluation, LREC-2004,* Lisbon, 26-28 May 2004.

Paice C.D. « Constructing literature abstracts by computer: techniques and prospects », *Information Processing & Management*, vol. 26, nº 2, 1990, p. 171-186.

Porter M.F. « An algorithm for suffix stripping », *Program*, vol. 14, 1980, p. 130-137.

Sampson G. *Empirical Linguistics*, London, Continuum, London, 2001.

Savoy J. « Recherche d'informations dans des corpus en langue française : Utilisation du référentiel Amarylis », *Technique et Science Informatiques,* vol. 21, nº 3, 2002, p. 345-373.

Strube M. & Hahn U. « Functional centering », *Proceedings* of *Association for Computational Linguistics*, Santa Cruz, 24-27 June 1996, Morgan Kaufmann, p. 270-277.

Véronis E., Véronis J. & Voisin N. *Les politiques mis au net*, Paris, Max Milo Editions, 2007.

Witten I.H., Gori M. & Numerico T. *Web Dragons*, Amsterdam, Elsevier, 2007.

Annexe : Liste de mots-outils ignorés en langue française

a	ces	est	n	s
à	cet	et	ni	se
ainsi	cette	été	on	soit
au	ci	il	ont	sont
aussi	comme	ils	or	sur
aux	d	l	ou	tous
avec	dans	la	p	tout
c	de	le	par	toute
car	des	les	plus	toutes
ce	donc	leur	pour	un
ceci	du	leurs	qu	une
cela	elle	mais	que	y
celle	en	même	qui	

Tableau A.1 *: Liste des 64 mots-outils éliminés avant de procéder à nos analyses*