Publié dans les
Actes 6ème Conférence en Recherche d'Information et Applications CORIA'09,
185-200, 2009
qui doivent être utilisés pour toute référence à ce travail

Indexation et représentation comparative : Application au discours électoral

Jacques Savoy

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

RESUME. *Cet article décrit quelques approches afin d'extraire les termes les plus représentatifs d'un site web ou d'un ensemble de documents en comparaison avec d'autres sites ou un corpus de référence. Nous montrons que la fréquence d'occurrence ou le rang des termes les plus fréquents peut fournir une première synthèse. Notre proposition s'appuie sur une distribution binomiale des mots et le calcul d'un score normalisé (score Z) mettant en lumière les termes comparativement les plus appropriés. Quelques exemples tirés des discours électoraux suisses ou français illustrent l'intérêt de l'approche suggérée.*

ABSTRACT. *This paper describes some possible approaches to automatic extraction of terms closely reflecting the content of a Web site or a set of documents by comparison of other sites or a given corpus. We show that the frequency of occurrences or the rank of the most frequent terms may provide a first overview. In the suggested method, we model 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 and French political speeches show the usefulness of the suggested method.*

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

KEY WORDS: *Summarization, Indexing, Probabilistic Word distribution, Discourse analysis.*

1. Introduction

Internet a mis à notre disposition un volume considérable d'information présentée sous divers formats (XML, HTML, pdf), médiums (texte, image, audio, vidéo) et couvrant tous les domaines de l'activité humaine. En recherche d'information (RI), on reconnaît volontiers que les moteurs de recherche ont joué un rôle de premier plan dans l'accroissement exponentiel du nombre de pages disponibles. Pour traiter automatiquement un tel volume d'information et face aux divers besoins des usagers, de nombreuses perspectives d'application s'ouvrent, domaines que l'on peut regrouper sous le terme de "fouille de textes" [KON 06].

Dans ce cadre, 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. Comme en RI, notre objectif consiste à indexer et à représenter d'une manière compacte une page, un site ou un ensemble de documents. 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 (ou des) date(s) antérieure(s).

Une telle représentation permet de 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 and Tracking*). La *blogosphère* [FAU 08] présente également un intérêt afin de suivre les sentiments ou opinions des internautes exprimés dans des billets d'information par rapport à un groupe de référence ou, dans une perspective dynamique, en mettant en lumière l'évolution temporelle des intérêts ou sentiments.

Dans cette communication, nous souhaitons explorer quelques pistes afin de définir et d'analyser diverses stratégies de représentation comparative. Nous nous limiterons cependant à des informations de nature textuelle pouvant correspondre à une page, à un site Internet ou à un ensemble de documents. Dans cette perspective, quelques travaux reliés à la génération automatique de résumé ou à l'extraction de termes significatifs seront présentés dans la deuxième section. Le domaine d'application, le discours électoral suisse, sera exposé dans la troisième section. Les diverses stratégies d'extraction seront décrites dans la quatrième section. Enfin, basé sur nos outils, une comparaison des discours électoraux en Suisse et une comparaison franco-suisse sera présentée dans la dernière section.

2. Extraction et résumé automatique

Dans la génération automatique de résumé, 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 [STU 96] (e.g., 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*. [GOL 99] distinguent entre deux types de résumé, 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 concordances [PAI 90]. 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 [KUP 95]. 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 [BER 00] 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 (et nécessitant 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). 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 [FUL 08].

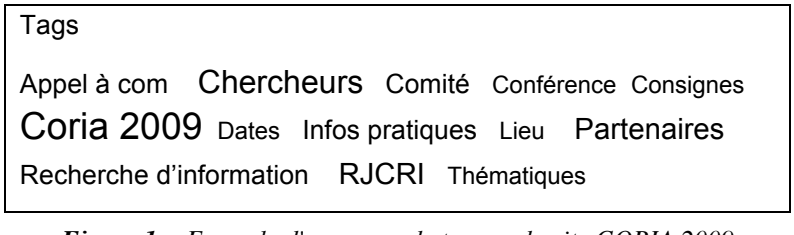

***Figure 1** : Exemple d'un nuage de termes du site CORIA 2009*

3. Champ d'étude et corpus d'évaluation

Dans l'exploration des diverses possibilités offertes afin de représenter de manière *comparative* un ensemble de documents, il nous est apparu important de pouvoir vérifier la qualité des résultats. Malheureusement, nous ne disposons pas

d'un corpus de test et d'une métrique adaptée. Notre démarche se voulant empirique, une justification expérimentale nous est alors parue importante afin de guider nos choix.

3.1. Justification du choix du corpus de référence

Afin de répondre à ces attentes, nous avons sélectionné les sites des partis politiques et, plus précisément, les discours ou programmes électoraux. Diverses raisons justifient notre choix. Premièrement, afin de garantir une grande homogénéité, nous avons 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 les formations entendent 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, ou des thèmes retenues. En campagne électorale, chaque 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 ses mots ou de ses formulations.

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 [LAB 07]. De plus, des études antérieures portant sur le discours politiques existent [LAB 03], [LAB 08] permettant une comparaison avec nos travaux. Signalons toutefois que ces derniers portent sur le discours gouvernemental et non électoral.

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, de syntaxe ou l'usage abusif d'abréviations.

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 ou des positions du parti dans cette perspective.

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. Enfin, nous avons indiqué le nombre de documents extraits de chaque site et son URL.

	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
Site internet	www.sp-ps.ch	www.pdc.ch	www.prd.ch	www.svp.ch

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

Nus avons limité notre analyse comparative aux quatre grands partis[1] 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 correspondants 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 %).

3.3. Prétraitement du corpus

Afin de définir une représentation comparative de chaque site (ou parti dans notre cas), nous avons choisi de tenir compte des termes présents ou absents. 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.

Notre système d'analyse ne fera pas de distinction entre majuscule et minuscule et les formes "Suisse" ou "suisse" seront considérées comme identiques. Certes, les

[1] 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.

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).

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”, “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 [SAV 02] supprimant la marque du pluriel (le ‘-s’ final ou la transformation de la séquence finale ‘-aux’ en ‘-al’). 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 de 15,4 %). Parfois la forme au singulier ou au pluriel s’avère aussi fréquente l’une que l’autre (e.g. 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 (e.g. 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 d’un vocabulaire sera sujet à des variations dues aux circonstances (le contexte, l’auditoire, intervention spontanée ou discours lue) 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 notre corpus. En effet, les textes proviennent de la même période, sont rédigés dans la même perspective et couvrent des objectifs très similaires.

4. Quelques stratégies d'extraction

L'indexation automatique en RI propose de définir l'importance de chaque terme des documents (et des requêtes) en tenant compte essentiellement de leur fréquence d'apparition (*tf*), de fréquence documentaire d'un terme (*df*, ou plus précisément de l’*idf* = log(*n*/*df*)) et de la longueur du document. Dans notre contexte, le nombre de documents ou de sites distincts demeure faible ($n = 4$ dans le cas présent) réduisant l'intérêt pour une mesure *idf*. En effet de nombreux mots apparaissent dans les quatre 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 ensemble de documents) en fonction d'autres sites (ou d'un corpus de référence).

4.1. Richesse lexicale et vocabulaire

Comme première approche, nous pourrions comparer la taille du vocabulaire utilisé dans les quatre sites. Dans cette perspective, nous pouvons estimer qu’un

lexique étendu correspond à un parti ayant de grandes ambitions, désirant couvrir tous les domaines [LAB 03]. A contrario, en présence d'un vocabulaire plus restreint, nous pourrions avancer l'hypothèse que 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 [BAA 01]. Comme l'indique le tableau 1, les quatre grands partis présentent des volumes assez différents. Comme le PRD propose le corpus le moins long, nous avons réduit les trois autres corpus à cette taille en ne retenant que les 26 639 premières formes (voir tableau 2).

	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 2** : Richesse lexicale en nombre de vocables (forme distincte) et nombre de vocables apparaissant qu'une seule fois (hapax)*

Si l'on compare, en prenant le même nombre de mots (soit 26 639), le nombre de formes différentes (vocables) utilisées par les quatre grands partis suisses, 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 serait 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 apparaissant qu'une seule fois (*hapax legomena*) contre 1 811 pour le PDC, 1 676 au PS et 1 511 au PRD. Ces informations nous fournissent un indice lexical global mais pas une représentation de la sémantique sous-jacente.

4.2. Fréquence d'occurrence

Afin de déterminer les mots nécessaires à refléter le 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, les, l, et, des, le, d, en, une, dans, du est, pour, que, un, etc.). Après élimination de 64 termes peu porteurs de sens, nous voyons mieux émerger les thèmes récurrents et communs à l'ensemble des formations et par différence, ceux propres à chaque parti.

Le tableau 3 indique pour chaque parti 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", "ne" et "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 3** : Les dix vocables les plus fréquents (et leur fréquence absolue) dans les discours des quatre partis*

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 le "droit", "étranger" et les problèmes de l'assurance-invalidité (AI).

Le recours à la fréquence d'occurrence permet de faire ressortir les vocables décrivant bien le contenu sémantique d'un document. De plus cette information peut définir l'importance de chaque terme dans une représentation (e.g., par des variations de fonte, taille ou style dans un interface de type "nuage de termes"). On prendra soin toutefois de normaliser cette fréquence lexicale en fonction de la longueur du document sous-jacent (e.g., *tf*/max *tf*).

Comme alternative, on peut retenir la fréquence lexicale uniquement comme clé de tri. L'importance de chaque terme se mesurerait alors en fonction uniquement 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[2]. Par exemple, la différence de rangs entre les vocables "pas" et "suisse" pour le parti UDC (voir tableau 3) 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

Si l'on désire connaître le vocabulaire spécifique à un parti, nous devons le comparer à une norme, par exemple à un corpus composé des sites des quatre partis. Comme le tableau 3 l'indique, plusieurs vocables apparaissent fréquemment dans plusieurs sites et ne permettent donc pas de discriminer de manière comparative les contenus. Avec un corpus de référence, nous pouvons observer quel vocable 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 [MUL 92].

Si nous regroupons tous les documents ou divers sites que l'on désire comparer, nous pouvons former un corpus **C**. Si nous désirons définir une représentation comparative par rapport à **C**, nous pouvons extraire les documents correspondant à un sous-ensemble noté **S**. Le reste du corpus sera noté **C-** (avec **C** = **S** ∪ **C-**).

Intéressons nous à un vocable particulier noté ω. Nous pouvons alors compter sa fréquence d'occurrence dans le sous-ensemble **S** (valeur notée *a* dans le tableau 4) et sa fréquence dans le reste du corpus **C-** (valeur notée *b*). Evidemment, la fréquence d'apparition dans tout le corpus **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 $n = a + b + c + d$.

	S	**C-**	
ω	a	b	$a + b$
non ω	c	d	$c + d$
	$a + c$	$b + d$	$n = a + b + c + d$

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

Sur la base des informations données dans le tableau 4, nous faisons l'hypothèse que le mot ω suit une distribution binomiale dans le sous-ensemble **S** avec comme

[2] La distribution de la fréquence d'occurrence suit une loi de puissance [BAA 01].

paramètres p et n'. Le paramètre p indiquant la probabilité d'apparition du terme ω qui peut être estimé par $(a+b)/n$ tandis que $n' = a+c$ correspond à la taille (en nombre de mots) du sous-ensemble **S**. Selon cette distribution binomiale, le nombre moyen d'occurrence du terme ω dans la partie **S** sera donc de $n' \cdot p$. Pour un vocable donné ω, le tableau 4 indique que le nombre observé dans le sous-ensemble **S** s'élève à a. Enfin, l'écart-type de la distribution binomiale s'estime par $n' \cdot p \cdot (1-p)$.

Sur ces éléments, nous pouvons calculer un score normalisé (ou score Z) pour chaque terme ω en tenant compte du nombre observé d'occurrences (valeur a) auquel on soustrait sa moyenne théorique (nombre prédit par la distribution binomiale) et que l'on divise par l'écart-type estimé.

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

La différence entre le nombre réellement observé a et la moyenne théorique ($n' p$) permet de déterminer les suremplois (différence positive) et les vocables sous-représentés (différence négative). Cette différence doit encore être divisée par un estimateur de l'écart-type pour retourner une valeur standardisée. Comme règle de décision, un score Z supérieur à 3 indiquera un suremploi significatif[3] du vocable dans le sous-corpus tandis que des valeurs négatives et inférieures à -3 signalent des sous-emplois marqués par rapport au corpus global.

5. Applications

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, nous avons décidé de comparer les discours électoraux tenus dans les deux pays. Dans ce but, nous avons récupéré 11 discours prononcés par S. Royal et 17 prononcés par N. Sarkozy[4] (section 5.2). Une différence de style 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, variation de forme qui peut influencer le choix du lexique.

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 5) ainsi que ceux qui sont sous-représentés (tableau 6). Pour établir ces listes ordonnées, nous

[3] 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.

[4] Ces deux sous-corpus représentent respectivement 93 479 mots pour S. Royal et 116 212 mots pour N. Sarkozy, soit une taille un plus élevée que celle des discours suisses (tableau 1).

avons calculé le score z normalisé de chaque terme, valeur indiquée à côté de chaque entrée dans les tableaux 5 et 6.

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

Les vocables apparaissant dans le tableau 5 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 3). 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 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].

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.

[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.

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 est 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 6). 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 l'un des thèmes majeurs concernait 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 6** : Les dix vocables les plus sous-employés dans les sites des partis suisses*

5.2. Comparaison Suisse - France

Entre les deux pays, les vocables fréquents montrent l'origine politique des documents mais des distinctions apparaissent également rapidement (voir tableau 7). Ainsi, on retrouve les formes "ne", "pas", "être" ou "nous" des deux côtés de la frontière, sans que ces éléments relèvent des informations très pertinentes. Nous aurions pu ajouter le vocable "politique" qui apparaît en onzième rang du côté français. De même, le recours différencié aux formes "suisse" et "france" ou "français" ne nous surprennent pas.

Parfois les petits mots font toute la différence et dans ce cas nous rencontrons un emploi marqué du "je" (ainsi que du vocable relié "j") dans les discours électoraux

français par rapport à ceux de la Suisse. Ce vocable "je" indique bien l'importance attachée à une personne, au chef du parti ou futur président dans l'Hexagone. Plus étonnant, la fréquence d'occurrence du pronom "je" s'avère statistiquement plus élevée pendant le deuxième tour de la campagne présidentielle que lors du premier tour [LAB 08b]. Le passage au second tour s'accompagne bien d'un changement au niveau lexical et cela se vérifie pour les deux candidats. Pour un chef de parti, 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". Ce vocable peut également s'expliquer, en partie, par la forme orale du discours électoral français.

	Suisse	**France**	**S. Royal**	**N. Sarkozy**
1	suisse	je	je	je
2	nous	pas	nous	pas
3	pas	ne	vous	ne
4	politique	nous	pas	france
5	être	france	france	nous
6	ne	vous	ne	veux
7	doit	veux	veux	si
8	droit	si	j	parce
9	notre	parce	notre	être
10	doivent	être	faire	français

***Tableau 7** : Les dix vocables les plus fréquents en Suisse et en France*

En complément, nous avons également calculé les quinze vocables sur-employés dans les discours des deux pays de même que les suremplois des deux candidats à l'Elysée (voir tableau 8). En premier lieu, on y retrouve les dénominations propres à chaque pays ("suisse", "france") de même que ceux rattachées à leurs institutions respectives ("canton", "conseil", "fédérale", "confédération", "UDC", "PDC" d'une part et, d'autre part, "présidentiel", "république"). De manière plus profonde, on retrouve, du côté français, les formes verbales "veux", "suis", "dire", "crois" (en 20e rang) ou la conjonction "parce que" indiquant un besoin explicatif indéniable, 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" (18e rang)) ainsi que les vocables "assurance" (17e rang), "économie" (24e rang) ou "culturelle" (27e rang).

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", forme apparaissant au 41e rang des

vocables les plus fréquents chez S. Royal et en 88e rang chez N. Sarkozy[6]. Ces résultats démontrent l'intérêt de la méthode proposée (tableau 8) en regard du recours à la fréquence d'occurrence absolue ou normalisé (tableau 7).

Suisse		**France**		**S. Royal**		**N. Sarkozy**	
Z	vocable	Z	vocable	Z	vocable	Z	vocable
28,1	suisse	39,8	je	29,1	vous	32,2	je
14,8	fédéral	26,4	france	23,7	je	25,4	pas
13,0	AI	23,0	vous	18,1	pacte	24,0	france
12,6	confédération	20,9	veux	13,8	sera	21,9	veux
12,5	UDC	17,3	français	13,7	salarié	20,8	français
12,4	étranger	16,8	parce	12,9	présidentiel	20,0	parce
11,6	franc	16,6	j	12,8	france	19,6	ne
11,4	doivent	16,2	pas	12,0	oui	15,9	république
10,7	canton	15,2	ai	11,9	femme	14,4	si
10,5	doit	13,1	république	11,6	logement	13,5	ai
10,4	neutralité	12,4	ne	11,1	construire	12,9	j
10,4	UE	12,3	suis	11,1	juste	12,5	ceux
10,3	conseil	12,0	dire	10,5	j	12,4	parler
10,0	fédérale	11,8	me	10,4	emploi	12,3	suis
9,8	PDC	11,6	ceux	10,3	nous	12,3	rien

***Tableau 8** : Les quinze vocables les plus surreprésentés de manière significative dans les deux pays*

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 correspondant. Les résultats permettent certes de se faire une idée du contenu d'un site, d'un document ou d'un ensemble de documents. 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.

Comme approche, nous proposons de recourir à un score normalisé (score z) sous l'hypothèse que la distribution des termes suit une loi binomiale. En appliquant cette méthode pour l'analyse comparative des discours électoraux en Suisse (élection

[6] Le vocable « europe » apparaît en 282e position des formes les plus fréquentes des discours politiques suisses.

d'octobre 2007), nous pouvons mettre en lumière les thèmes porteurs et caractéristiques des quatre grandes formations suisses. Il en ressort clairement que le parti de la droite dure et populiste (UDC) situe son débat autour de la peur de "l'étranger", du refus de toute "naturalisation" facilitée, des "rentes" de l'assurance-invalidité "AI", ainsi qu'une volonté de maintenir une vision très stricte de la "neutralité". Le parti PRD de la droite modérée axe sa thématique sur la réforme fiscale tandis que le parti du centre (PDC) garde son orientation "enfant" et "famille". La gauche tient quelques positions classiques ("autogestion" ou "sociale") mais s'ouvre vers les problèmes de l'environnement (taxe sur le "CO2"), thème central de leur alliée, le parti écologiste.

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. Le discours helvétique met un accent plus important sur les formes verbales "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" ou "pacte présidentiel") 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 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).

Bibliographie

[BAA 01] Baayen, H.R. "*Word Frequency Distributions*", Kluwer Academic Publishers, Dordrecht, 2001.

[BER 00] Berger, A.L. & Mittal, V.O. "OCELOT: A system for summarizing web pages", Proceedings *ACM-SIGIR'2000*, p. 144-151.

[FAU 08] Fautsch, C. & Savoy, J. "Stratégies de recherche dans la blogosphère", *Document Numérique*, vol. 11, n° 1-2, 2008, p. 109-132.

[FUL 08] 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, 2008.

[GOL 99] Goldstein, J., Kantrowitz, M., Mittal, V. & Carbonell, J. "Summarizing text documents", Proceedings *ACM-SIGIR'99*, p. 121-128.

[KIL 07] Kilgarriff, A. "Googleology is bad science", *Computational Linguistics*, vol. 33, nº 1, 1991, p. 147-151.

[KON 06] Konchady, M. "*Text Mining Application Programming*", Ch. River, Boston, 2006.

[KUP 95] Kupiec, J., Pedersen, J. & Chen, F. "A trainable document summarizer", *Proceedings ACM-SIGIR'95*, p. 68-73.

[LAB 03] Labbé, D. & Monière, D. "*Le discours gouvernemental. Canada, Québec, France (1945-2000)*", Champion, Paris, 2003.

[LAB 07] Labbé, C. & Labbé, D. "Baudelaire, Rimbaud et Verlaine", Actes *Aspects linguistiques du texte poétique*, Brest, 2007.

[LAB 08] Labbé, D. & Monière, D. "*Les mots qui nous gouvernent. Le discours des premiers ministres québécois : 1960-2005*", Monière-Wollank, Montréal., 2008.

[LAB 08b] Labbé, D. & Monière, D. "Je est-il un autre ?", Actes *JADT 2008* (Journées internationales d'Analyse statistique des Données Textuelles), 2008, p. 647-656.

[MAN 99] Mani, I. & Maybury, M.T. "*Advances in Automatic Text Summarization*", The MIT Press, Cambridge, 1999.

[MUL 92] Muller, C. "*Principes et méthodes de statistique lexicale*", Honoré Champion, Paris, 1992.

[PAI 90] Paice, C.D. "Constructing literature abstracts by computer: techniques and prospects", *Information Processing & Management*, vol. 26, nº 2, 1990, p. 171-186.

[SAV 02] Savoy, J. "Recherche d'informations dans des corpus en langue française : Utilisation du référentiel Amarylis", *TSI,* vol. 21, nº 3, 2002, p. 345-373.

[STU 96] Strube, M. & Hahn, U. "Functional centering", Proceedings of *Association for Computational Linguistics*, Morgan Kaufmann, 1996, p. 270-277.

Annexe : Liste de mots-outils ignorés

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*