Lexical Analysis of Obama's and McCain's Speeches

Jacques Savoy
Computer Science Dept., University of Neuchatel
Rue Emile Argand 11, 2009 Neuchatel (Switzerland)
Jacques.Savoy@uninе.ch

Abstract

This paper describes a US political corpus comprising 189 speeches given by senators John McCain and Barack Obama during the years 2007-08. We presented the main characteristics of this collection and compare the common English words most frequently used by these political leaders with ordinary usage (Brown corpus). We then discuss and compare certain metrics capable of extracting terms best characterizing a given subset of the entire text corpus. Terms overused and underused by both candidates during the last US presidential election are determined and analyzed from both a statistical and dynamic perspectives.

1 Introduction

The presidential election was the major political event in the United States in 2008. During this campaign the candidates (or their speechwriters) wrote various speeches that would hopefully convince undecided voters, to encourage their supporters and to make obvious that they were the best candidate for the job. The words and expressions used in their discourses were therefore not chosen randomly but rather to reflect these various objectives. Since the candidates' speeches targeted the same election, they expressed their views during the same period and concerned the same goals and related topics, we were thus able to compare the speeches more objectively than say various literary works selected from different periods, styles (e.g., tragedies, novels) and genres (prose vs. poetry). We must however recognize that in politics the official version is usually the spoken one. But we can consider that the written version usually available on each candidate's web site reveals accurately the speaker's real intent. Also, these freely available texts usually contain few spelling errors and abbreviations, which from the information technology point of view render their use without real problems. Finally, from the perspective of interpreting and verifying results, we deem it easier to work with political speeches rather than with texts from more technical domains.

Using words extracted from these speeches, our objective is to define the various terms that can characterize well each subset of our overall US political corpus. These subsets could be defined according to date (2007 vs. 2008), author (J. McCain vs. B. Obama), topic (e.g., energy vs. foreign policy), form (spoken vs. written), or target audience (e.g., business vs. academic). For the purposes of this paper, we limit ourselves to only distinguishing the author and the date (month and year).

The rest of this paper is organized as follows. Section 2 presents a brief overview of related work in political discourse analyses. Section 3 provides an overview of our US political corpus while Section 4 discusses certain metrics used to define and weight the terms best characterizing the differences between two (or more) sets of documents (corpus partitions). Section 5 describes the main differences revealed through comparing the two candidates, while Section 6 shows their differences from a dynamic perspective. Section 7 displays how we follow the importance of a given topic throughout the entire campaign, on a month-by-month basis. Finally Section 8 presents some additional comparisons elements such as sentence lengths and distribution of POS across candidates.

2 Related Work

In our analysis of political corpora and lexical analysis, we pay tribute to the work done by Labbé & Monière (2003) in comparing the three sources of government speeches (e.g., speeches

from the Throne (Canada), inaugural speeches (Quebec) and investiture speeches (France)). The advantage of their work is that it covers documents written in the French language, over a relatively long period of time (50 years, from 1945 to 2000) and makes it possible to compare political discourses from these countries. This corpus however only consists of government speeches, and thus they were not necessarily written for electoral purposes. We can expect certain differences between a Prime Minister in charge of a government and one who is hoping to be elected (Herman, 1974). Even though these government speeches express the ideas of distinct political parties, according to Labbé & Monière (2003) they tended to be more similar than expected, mainly due to institutional constraints. As such, continuity clearly imposes stronger constraints than political cleavages. They did note however a certain trend towards longer speeches (perhaps related to television broadcasting and the complexity of the underlying questions).

Measuring lexical richness objectively is a complex problem especially given that a well-grounded operational definition does not exist. To do so we need to take into account the number of distinct words, vocabulary diversity and expansion over time, lexical specificity, etc. (Baayen 2008). According to Labbé & Monière (2003), the reason for vocabulary increases cannot be attributed to a single and well-defined event, but rather may occur when a strong personality takes power, such as that of Prime Minister Trudeau (1968-72) in Canada, or Rocard (1988) and Bérégovoy (1992) in France.

There are of course other pertinent research questions related to our research. One might wish to discover the name of the actual speechwriter behind each discourse (as, for example, T. Sorensen behind President Kennedy (Carpenter & Seltzer, 1970)). We might also compute textual distances between speeches, sets of speeches or political leaders (based on their speeches) to measure the relative distance between them (Labbé, 2007). Based on this information, we could then draw a political map showing the various political leaders according to their respective similarities (Labbé & Monière, 2003).

3 Our US Political Corpus

This corpus contains speeches we downloaded from the two candidates' official web sites. For each speech, we added a few meta-tags to store document information (e.g., document identifier, date, location, title), and we also cleaned them up by replacing certain UTF-8 coding system punctuation marks with their corresponding ASCII code symbol. This involved replacing single (‘’) or double quotation marks (“”), with the (') or (") symbols, and the removal of diacritics found in some certain words (e.g., “naïve”). To improve matching between surface forms we also replaced uppercase letters by their corresponding lowercase, except for those words written only with capital letters (e.g., “US,” “FEMA” (Federal Emergency Management Agency)).

On the other hand, we did not try to normalize various word forms referring to the same entity such as “US,” “United States,” “United States of America,” or “USA” (“America,” “our country” etc.). We assume that the authors maintain the same form across the two years and that they will use the same spelling. This assumption is reasonable, given that both candidates would follow the same objectives and their speeches would be extracted from the same time period.

February 10th, 2007: Senator Barak Obama (IL) announced his candidacy for President
April 25th, 2007: Senator John McCain (AZ) announced his intention to run for President
February 5th, 2008: Super Tuesday
June 7th, 2008: Hillary ended her campaign
August 23th, 2008: John Biden nominee as vice-president (D) August 25th-28th, 2008: Democrat convention
August 30th, 2008: Sarah Palin nominee as Vice President (R) September 1st-4th, 2008: Republican convention
September 1st, 2008: Official campaign starts
November 4th, 2008: Election day
January 20th, 2009: Inauguration Day

Table 1: Main events during the latest US presidential campaign

3.1 Overall Statistics

Obama's speeches were downloaded from `www.barackobama.com`, beginning with the first on February 10, 2007 and ending with that on September 18, 2008 (Table 1 indicated the main dates of this election). In total our corpus contains 114 speeches (37 in 2007, 77 in 2008), for a total data size of 1.76 Mb (0.7 Mb for 2007, 1.06 Mb for 2008). For the Republican Party's speeches, we downloaded them from `www.johnmccain.com` beginning on April 25th, 2007. This second subset contains 75

speeches (23 for 2007, 52 for 2008), for a total of 1.04 Mbytes (0.32 Mb for 2007, 0.72 Mb for 2008).

	McCain	Obama
2007	23	37
01/2008	3	7
02/2008	2	6
03/2008	3	6
04/2008	12	9
05/2008	10	9
06/2008	10	12
07/2008	7	14
08/2008	4	9
09/2008	1	5
Total	**75**	**114**

Table 2: Distribution of speeches by date and author

The data listed in Table 2 shows see that McCain gave fewer speeches than Obama (75 vs. 114). Their distribution across the entire period shows that Obama tended to give more speeches, except for the months of April and May, 2008.

From inspecting the number of word tokens per author and date (see Table 3), we see that B. Obama reduced the volume of his speeches over the last year (2007 mean: 3,402, 2008 mean: 2,457), and that they tended to have the same mean length as McCain's speeches (2,349), who showed a more stable mean across the two years (computation done with **R** (Crawley, 2007) and text processing with Perl (Nugues, 2006)).

	McCain	Obama
Total tokens	176,457	315,043
in 2007	54,319	125,857
in 2008	122,138	189,186
Tokens/speech	2,353	2,764
in 2007	2,362	3,402
in 2008	2,349	2,457
Number forms	8,715	9,071
in 2007	5,108	6,547
hapax in2007	2,171	2,476
freq. ≤ 4 in 2007	3,699	4,411
in 2008	7,410	7,169
hapax in 2008	2,866	2,552
freq. ≤ 4 in 2008	5,010	4,557

Table 3: Statistics on speeches, listed by year and author

Table 3 shows also the number of distinct word forms (or vocabulary size) used by each candidate. It is interesting to note that of the 7,410 distinct word forms that McCain used in his speeches in 2008, 2,866 (or 38.7%) word forms were used only once (a phenomenon known as *hapax*). Words used four times or less represent a rather large proportion, namely 67.6% of the total (or 5,010 forms). An analysis of Obama's vocabulary reveals a similar pattern. Also noteworthy is that even though McCain gave fewer speeches than Obama in 2008, his vocabulary tended to be larger (7,410 vs. 7,169).

3.2 Most Frequent Words

Next we compared the vocabulary found in our US political corpus with that of other written English text formats. Table 4 lists the 20 most frequent lemmas (e.g., the lemma "be" includes the forms "be," "is," "are," "was," etc.) extracted from the Brown corpus (reflecting common American usage in the early 60s) and compares them with those of our US political corpus, through applying the Stanford POS tagger system (Toutanova & Manning, 2000). There is of course a time gap but given the forms shown in Table 4, this does not seem to play a really significant role and would thus not invalidate any comparisons.

	Brown		US	
Rank	Lemma	Freq.	Lemma	Freq.
1	the	6.90%	the	4.77%
2	be	3.86%	be	3.80%
3	of	3.59%	and	3.79%
4	and	2.85%	to	3.32%
5	to	2.58%	of	2.67%
6	a	2.28%	that	2.18%
7	in	2.06%	a	1.98%
8	he	1.92%	in	1.89%
9	have	1.23%	we	1.81%
10	it	1.08%	I	1.48%
11	that	1.05%	have	1.35%
12	for	0.89%	for	1.17%
13	not	0.87%	not	1.17%
14	I	0.83%	our	1.12%
15	they	0.82%	it	1.03%
16	with	0.72%	will	0.94%
17	on	0.61%	this	0.81%
18	she	0.60%	do	0.66%
19	as	0.59%	you	0.64%
20	at	0.53%	on	0.61%

Table 4: Top 20 word forms found most frequently in Brown and US corpus

From Table 4 it can be seen that "the" tends to occur more frequently in ordinary language (6.9%) than in political speeches (4.77%). What is more interesting is the conjunction "that" which ranks 6th in our US political speeches but only 11th in the Brown corpus. This indicates that politicians tend to produce longer sentences with more complex syntax, reflecting a need to

be more precise or to explain certain problems in depth. Political speeches are often characterized by the frequent use of the pronoun "we" (ranked 9th compared to 23rd in the Brown corpus). The verb "will" shows a similar pattern (16th vs. 35th in the Brown corpus). The pronoun "he" however (8th in the Brown corpus) is used less in our US corpus, where it is ranked 45th. The difference is even greater for the pronoun "she" (18th vs. 221th). Applying the Wilcoxon matched-pairs signed-ranks test (Conover, 1971), we can verify if both rankings reflect a similar words usage. In the current case, we must reject this hypothesis (p-value < 0.001).

4 Metrics

These findings may be used to distinguish between speeches given for political reasons and in comprising ordinary language. Our goal however is to design a method capable of selecting terms that clearly belong to one type of document and that can be used to properly characterize it (Daille, 1995), (Kilgarriff, 2001). Various authors have suggested formulas that could meet this objective, and they are usually based on a contingency table such as that shown below.

	S	C-	
ω	a	b	$a+b$
not ω	c	d	$c+d$
	$a+c$	$b+d$	$n=a+b+c+d$

Table 5: Example of a contingency table

The letter a represents the number of occurrences (tokens) of the word ω in the document set S (corresponding to a subset of the larger corpus C). The letter b denotes the number of tokens of the same word ω in the rest of the corpus (denoted C-) while $a+b$ is the total number of occurrences in the entire corpus. Similarly, $a+c$ denotes the total number of tokens in S. The entire corpus C corresponds to the union of the subset S and C- (C = S∪C-), and contains n tokens ($n=a+b+c+d$).

Based on the MLE (Maximum Likelihood Estimation) principle the values shown in a contingency table could be used to estimate various probabilities. For example we might calculate the probability of the occurrence of the word ω in the entire corpus C as Prob(ω) = $(a+b)/n$ or the probability of finding in C a word belonging to the set S as Prob(S) = $(a+c)/n$.

As a first approach in determining whether a given word ω could be used to describe the subset S quite adequately, we might consider two events. First we could estimate the probability of selecting the word ω in the entire corpus C (Prob(ω) = $(a+b)/n$). On the other hand, the probability of selecting a word in C belonging to the set S could be estimated by Prob(S) = $(a+c)/n$. Then if we consider selecting from C an occurrence of the word ω belonging to the set S, we could estimate this probability using Prob(ω ∩ S) = a/n. However we could also assume that the joint event (ω ∩ S) would be independent (by chance only) of both events (ω and S), which in turn would lead to another estimate, Prob(ω) · Prob(S).

To comparing these two estimates we would use the approach adopted by the mutual information (MI) measure (Church & Hanks, 1990), defined as:

$$I(\omega;\mathrm{S}) = \log_2\left[\frac{\mathrm{Prob}(\omega\cap\mathrm{S})}{\mathrm{Prob}(\omega)\bullet\mathrm{Prob}(\mathrm{S})}\right]$$

$$= \log_2\left[\frac{a}{(a+b)}\bullet\frac{n}{(a+c)}\right] \qquad (1)$$

When the two estimates are close (I(ω;S) ≈ 0), this means there is no real association between the word ω and the set S. In such cases, the occurrences of word ω in S can be explained simply by chance. When the word ω is used more often within S, then a positive association develops between them and we could find that Prob(ω ∩ S) > Prob(ω) · Prob(S), resulting in I(ω;S) > 0. Finally, if Prob(ω ∩ S) << Prob(ω) · Prob(S), this indicates that the two events are complementary and thus I(ω;S) < 0.

	Obama'08	**US**	
zionist	1	0	1
without	189,185	302,314	491,499
	189,186	302,314	491,500

Table 6: Distribution of the word "zionist" in Obama and US Speeches

Table 6 illustrates how the word "zionist" is distributed in Obama's speeches in 2008 and in the rest of our US corpus. The resulting MI measure is I("zionist"; Obama'08) = 1.38, indicating an association between the two events (this value is in fact the largest among the MI values, as shown in Figure 1). In our example the word "zionist" occurs just once in one

Obama's speech in 2008. According to our MI measure, this rare event returns a high MI value, tending to indicate a real association between the word "zionist" and Obama's vocabulary. Only one occurrence of this term can be found however and to ignore this particular case, it is suggested that the additional constraint of $a \geq 5$ be imposed.

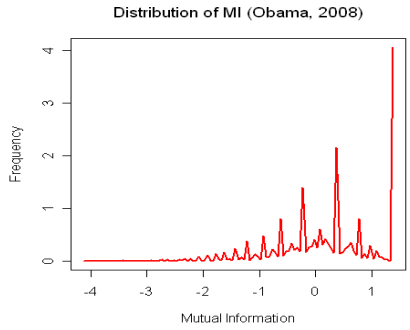

Figure 1: Distribution of Mutual Information values (Obama, 2008)

The chi-square (χ^2) measure (Manning & Schütze, 2000) provides a second approach to measuring the association between a word and a set of documents. This method allows us to compare the observed frequency (e.g., the value *a*) with the expected number of tokens, under the assumption that the two events (ω and S) are independent. This latter value is estimated using as $n \cdot \text{Prob}(\omega) \cdot \text{Prob}(S) = n \cdot (a+b)/n \cdot (a+c)/n = (a+b) \cdot (a+c)/n$. Rather than being limited to comparing the single cell storing the value *a*, we repeat this for the other three cells, namely *b*, *c*, and *d*.

Equation 2 below shows the general formula used to compute the chi-square measure, where o_{ij} indicates the observed frequencies and e_{ij} the expected frequency stored in cell ij.

$$\chi^2 = \sum_{i,j=1,2} (o_{ij} - e_{ij})^2 / e_{ij} \qquad (2)$$

According to the independence hypothesis, the χ^2 distribution follows a chi-square pattern, with 1 degree of freedom (dof). In order to infer valid conclusions, we usually add the constraint that each cell must have at least a minimal frequency (e.g., $o_{ij} \geq 5$). This results in a major reduction in the terms being analyzed, from 7,410 to 2,131 (28.8%) for McCain in 2008, and from 7,169 to 2,306 (or 32.2%) for Obama (see Table 3).

	McCain'08	**US-**	
McCain	19	360	379
without	122,119	369,002	491,121
	122,138	369,362	491,500

Table 7: Distribution of the word "McCain" in the McCain and US Speeches

As shown in Table 7, the word "McCain" is distributed throughout McCain's speeches in 2008 and in the rest of our corpus. This word occurs 19 times in the subset and the expected frequency, under the assumption of independence, is 94.18. The difference for this cell is rather large (19 - 94.18), and the resulting χ^2 is also quite high 80.46. Comparing this value with the limit value 6.63 (α = 0.01, 1 dof, or 10.83 with α = 0.001), we can reject the hypothesis that the word "McCain" is distributed randomly between the two disjoint sets of our US political corpus. In fact this term is used less by McCain than the other speaker (e.g. the Senator McCain does not introduce himself as "McCain said …"). This method owns the advantage of having a clear decision rule. We must however ignore a large set of words (around 70%, see Table 3) that occur fewer than 5 times in a sub-corpus.

As a third approach, we could measure the association between a given word and a corpus through computing the log-likelihood value (denoted G^2), see (Dunning, 1993), (Manning & Schütze, 2000). This method could be appealing when faced with relatively low frequency values (e.g., less than 5) because such events are also important in describing various linguistics phenomena. Based on our notation, the G^2 measure is defined in Equation 3 (Daille, 1995).

$$\begin{aligned} G^2 = 2 \cdot [& a \cdot \log(a) + b \cdot \log(b) + c \cdot \log(c) + d \cdot \log(d) \\ & - (a+b) \cdot \log(a+b) - (a+c) \cdot \log(a+c) \\ & - (b+d) \cdot \log(b+d) - (c+d) \cdot \log(c+d) \\ & + (a+b+c+d) \cdot \log(a+b+d+c)] \end{aligned} \qquad (3)$$

	Obama'08	**US-**	
the	8,756	15,081	23,837
Without	180,430	287,233	467,663
	189,186	302,314	491,500

Table 8: Distribution of the Word "*the*" in the Obama and US Speeches

We applied this measure in our corpus and Table 8 shows an example (the word "the" in Obama's 2008 speeches). The resulting G^2 value

is 32.92, a relatively high value. This thus tends to indicate a significant association between the determinant "the" and Obama's speeches, at least for those given in 2008. This method does not however provide any direct indication that the word tends to be over or underused (which is the case here).

Finally, we suggest using Muller's approach (Muller, 1992) to obtain a Z score for each term. To do so we apply Equation 4 to standardize the underlying random variable, removing the mean (centered) and dividing it by its standard deviation (reduced). The resulting Z score value is also known as the standard score.

$$\text{Z score}(\omega) = \left[\frac{a - n' \bullet \text{Prob}(\omega)}{\sqrt{n' \bullet \text{Prob}(\omega) \bullet (1 - \text{P rob}(\omega))}} \right] \quad (4)$$

In Equation 4 we assume that the word ω has a binomial distribution with parameter p and n'. The parameter p could be estimated (MLE) as $(a+b)/n$ with $n' = a+c$ corresponding to the size of the set S (see Table 5). In our opinion however the word distributions resembles the LNRE distributions (Large Number of Rare Events (Baayen, 2001)), and we would therefore suggest smoothing the estimation of the underlying probability p as $(a+\lambda)/(n+\lambda \cdot |V|)$, where λ is a parameter (set to 0.5 in our case) and |V| indicates vocabulary size. This method called Lidstone's rule is a generalization of Laplace's method (in which λ is fixed at 1) (Nugues, 2006). This modification will slightly shift the probability density function's mass towards unseen words (or words that do not yet occur) (Manning & Schütze, 2000).

As a rule governing our decision we would consider those terms having a Z score between -2 and 2 as words belonging to a common vocabulary, as compared to the reference corpus (e.g. "might," "road" or "land" in our case). A word having a Z score > 2 would be considered as overused, while a Z score < -2 would be interpreted as an underused term. The threshold limit of 2 corresponds to the limit of the standard normal distribution, allowing us to only find 5% of the observations (around 2.5% less than -2 and 2.5% greater than 2).

The empirical distribution of the Z score values is displayed in Figure 2 where the limit of 2 is represented by two straight lines and the limit of 2.5% of the observations by dotted lines. This figure shows that we have slightly more than 2.5% of the observation having a value greater than 2 (precisely 3.25%) or lower than -2 (3.5% for the current distribution).

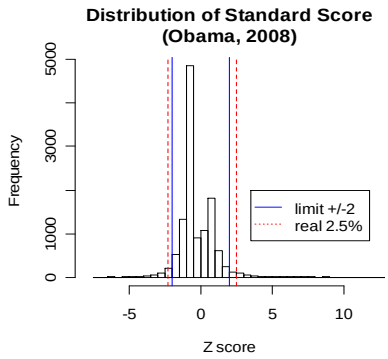

Figure **2**: Distribution of the Z score values (Obama)

From applying this computation to the word "zionist" (Table 6), the resulting Z score is 0.99, indicating a term that cannot be considered as overused. From Table 7, the word "McCain" has a Z score of -7.75 clearly indicating that it is a word underused by Senator J. McCain.

5 Differences Between Authors

We applied our Z score to specifically determine which terms each of the two political leaders used more (Z score > 2) and also to separate them from the more common political vocabulary (-2 ≤ Z score ≤ 2). It is however important to specify which corpus was used as reference. To do so we could compare the speeches given by Obama in 2008 with the entire US corpus (to see how his terms differ from those used by McCain) or with only those speeches given by the same speaker (to verify how the author's vocabulary varies throughout the campaign).

Table 9 lists the top overused and underused terms for both candidates, compared to the entire US political corpus. We examined all speeches (e.g., labeled "McCain") or only those speeches given in a specified year (e.g., only 2008 labeled "McCain … 2008"). As the table shows, terms usually overused by one candidate tend to appear as underused by the other. For example, the conjunction "because" and the adverb "why" are overused by Obama, reflecting his intention to explain the situation. He also overuses the name "Bush" and "McCain" (as shown in Section 4).

When considering the whole year 2008 month per month, we can find that Obama tends to overuse the term "that" and to underuse the word "Obama". For McCain, no word can be defined as overused during all months, and only the verb

form "is" is underused during the different months of year 2008.

A comparison of 2007 and 2008 demonstrates there is shift towards more political or electoral content in 2008 ("jobs," "government" or the other candidate's name).

	Overused	Underused
McCain	government, Obama, honor, freedom, power, public, …	because, why, McCain, Bush, street, working, …
2007	property, freedom, Islamic, construe, Reagan, enemy, …	because, school, jobs, McCain, children, working, …
2008	Obama, government, Canada, federal, small, judicial, …	why, because, McCain, college, Bush, …
Obama	because, why, McCain, college, Bush, street, …	government, Obama, honor, freedom, intend, …
2007	bullet, page, Joshua, Chicago, kids, poverty, …	senator, economic, tax, John, trade, government, …
2008	McCain, John, Bush, jobs, Washington, …	government, Obama, Congress, public, law, …

Table 9: Terms overused and underused in speeches by Obama and McCain when compared with the entire corpus

6 Dynamic Analysis

To provide a second perspective, we examined the speeches given by one candidate (Obama in our case) during the 2008 and on a month-by-month basis (arbitrary subdivision). Table 10 shows this comparison for the entire US corpus and Table 11 lists all speeches delivered by the same speaker.

2008	Overused	Underused
Jan.	deficit, Kennedy, Caroline, …	government, energy, oil, McCain,
Feb.	Orleans, NAFTA, FEMA, …	oil, power, nuclear, security, …
Mar.	regulator, Wright, black, …	energy, worker, oil, tax, …
Apr.	union, labor, worker, …	war, nuclear, government, …
May	Ryan, manufacturing, heroes, …	nuclear, market, Iraq, …
Jun.	Israel, patriotism, cities, …	politics, market, war, veteran, …
Jul.	Berlin, women, cyber, …	politics, insurance, cost, Israel, …
Aug.	Joe Biden, McCain, oil, …	war, reform, law, …
Sep.	financial, school, regulator, …	war, Iraq, oil, …

Table 10: Terms overused and underused in Obama's speeches when compared to the entire US corpus

2008	Overused	Underused
Jan.	deficit, Kennedy, assumption, …	McCain, million, energy, oil, …
Feb.	Orleans, NAFTA, FEMA, gulf, …	world, oil, women, history, …
Mar.	regulatory, Wright, black, war, …	you, energy, worker, tax, …
Apr.	labor, worker, union, trade, …	war, school, education, …
May	hemisphere, Cuba, Latin, freedom, …	Iraq, kids, nuclear, market, …
Jun.	Israel, patriotism, Jewish, cities, …	politics, war, veteran, people, …
Jul.	Berlin, women, cyber, Marshall, …	politics, change, tell, story, …
Aug.	Joe Biden, oil, energy, renewable, …	war, white, school, law, …
Sep.	financial, school, courses, McCain,…	war, Iraq, oil, energy, women, …

Table 11: Terms overused and underused used by Obama in selected monthly speeches when compared to all his speeches

The contents of the two tables are fairly similar, revealing very little impact, regardless of whether we compared speeches to the entire US corpus or only those given by Obama. An analysis of the terms overused for some months shows that Obama tends to present his patriotism ("patriotism" in June in response to McCain's attacks), his travels to Europe ("Berlin" in July), his selection for Vice President and the impact of oil prices ("Joe Biden," "oil," "renewable," in August) or the financial crisis ("financial," "regulator" in September). During 2008 he also uses more traditional topics such as Pastor "Wright" in March, "union," "labor," "worker," in April, or problems with "cities" in June. By contrast, during the months of April, May, August and September, the war in Iraq was clearly not a recurrent topic ("Iraq" was underused).

7 Thematic Follow-up

The Z score value associated with a word could also be used to reveal the evolution of a given topic during a specific time period, which in our case was 2008. This value was computed for each candidate and then compared with the entire US corpus. Through applying the same limits to the Z score, we could define overuse, underuse or normal use of specific terms during a given month.

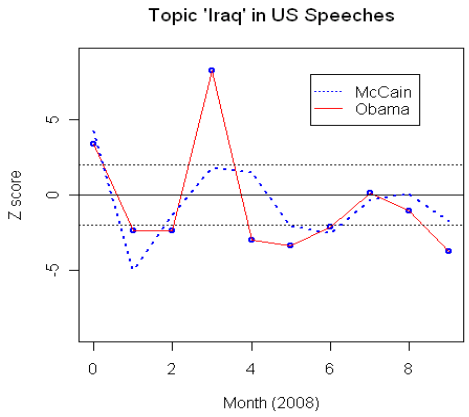

Figure 2: Z score value for "Iraq" topic variations

The Z score associated with the word "Iraq" changed for both candidates during the year 2008, as shown in Figure 2. The first value (x=0) shows the Z score throughout 2007, and we also see that while his issue was clearly present during 2007, during the first two months on 2008 it tended to decline. Obama frequently reintroduces this term in March, while McCain does so in March and April. Subsequently the topic tends only to be mentioned with only average frequency, while in September it tends to totally disappear from the campaign debate.

Clearly, as shown above in Figure 3, the term "jobs" is underused in 2007 by both candidates, while Obama reintroduced this question in the presidential campaign during February, and used it intensively in April and June. McCain ignored this topic until July when he overused the term. He then frequently reintroduced this word and during September 2008 both candidates tended to employ the word with average frequency.

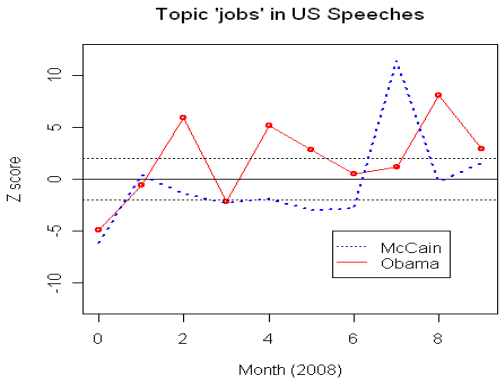

Figure 3: Z score value variations for the topic "jobs"

8 Additional Comparisons

While our comparative study was mainly based on single words, we could also consider additional speaker characterization features such as the length of their sentences (see Section 8.1) or the distribution of various parts of speech (POS) used (Section 8.2).

8.1 Sentence Length

In order to distinguish speeches made by two different politician leaders, we could consider sentence length (number of words). For McCain, the mean number of words per sentence is 25.46 (median: 23, std: 15.51; min: 1; max: 393, sample size: 9,702), and this mean value is fairly stable across the two years (in 2007, mean: 26.18; in 2008: mean: 25.19). Applying a *t*-test, we can see that the mean difference between the two years is statistically significant (p-value<0.01).

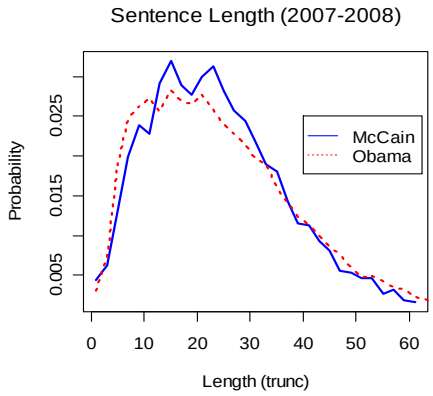

Figure 4: Sentence length distribution

For Obama, the mean sentence length is 26.05 (median: 23, std: 16.45; min: 1; max: 152, sample size: 17,804), and this value is fairly stable across the two years (in 2007, mean: 24.73; in 2008: mean: 26.78). Comparing the mean between the two years, a *t*-test indicates that the mean difference is statistically significant (p-value<0.001). As shown in Figure 4, the overall distribution for the two speakers is quite similar.

8.2 POS Distribution

To distinguish between texts written by two (or more) authors, we could analyze the frequency of most important POS (namely, noun, verb, adjective or adverb). We could also consider other POS (determinants, pronoun, conjunctions and prepositions), or related items such as dollar signs and numbers. To do so we could use the Stanford POS tagger system (Toutanova & Man-

ning, 2000), which automatically assigns the corresponding POS to each word.

Based on our own observations of the main differences between the two presidential candidates, we found the greatest differences were in the distribution of nouns, adjectives and verbs (without modal forms). The data in Table 12 shows that McCain used 38,442 nouns in his speeches (year 2007 & 2008). Given that the volume of his speeches represents only 34.8% of the total, the expected number of nouns in McCain's speeches would be $0.348 \cdot 100{,}238 = 33{,}456$.

	McCain	**Obama**	Total
noun	38,442	61,796	100,238
adj	13,494	20,158	33,652
verb	24,647	58,101	82,748
adv	7,984	18,128	26,112
Total	84,567	158,183	242,750
%	34.8%	65.2%	100%

Table 12: Distribution of main POS tags by politician

According to data depicted in Table 12, McCain tends to overuse nouns and adjectives, and thus his style seems to be more descriptive. This trend towards nouns in his campaign is also reflected by the use of buzzwords ("Country first: Reform, prosperity, peace"). On the other hand, Obama uses more active speech, preferring verbs such as "need" and "believe". Using the χ^2 test (Conover, 1971), we can infer that both styles are statistically different (p-value < 0.001). We also noted that he uses more frequent determinants, prepositions, as well as dollar signs and numbers, thus indicating a need to quantify his discussions.

9 Conclusion

In this paper we described the elaboration of a political corpus comprising 189 electoral speeches given by senators J. McCain and B. Obama. We suggested using a Z score combined with a smoothing technique of the underlying probability to identify those terms that adequately characterize subsets of this corpus and then we compared this measure with mutual information, chi-square and log-likelihood approaches. Through applying this Z score method to various corpus subsections we showed the most significant words used by both candidates during the two years. We also demonstrated how we can track the most overused and underused terms used by a given speaker or the how the treatment of a given topic varied during the campaign.

This study was limited to single words but in further research we could easily consider longer word sequences. Important trigrams associated with McCain could be for example: "health care system," "foreign oil dependence" while for Obama we found "million new jobs," "we can choose."

Other sources of information could be used to characterize and complement our electoral speeches analyses, such as the speech version actually delivered (including characteristics as intonation, prosody, stops and speaker indecision) to identify when the speaker is really at ease or unconformable with a given topic.

Acknowledgments

This research was supported in part by the Swiss National Science Foundation under Grant #200021-113273.

References

Harald R. Baayen. 2001. *Word Frequency Distributions*. Kluwer Academic Press, Dordrecht, NL.

Harald R. Baayen. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press, Cambridge, UK.

Ronald H. Carpenter and Robert V. Seltzer 1970. On Nixon's Kennedy Style. *Speaker and Gavel*, 7:41.

Kenneth W. Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22-29.

W. J. Conover. 1971. *Practical nonparametric Statistics*. John Wiley & Sons, 2nd Ed., New York.

Michael J. Crawley. 2007. *The **R** Book*. John Wiley & Sons, London.

Béatrice Daille. 1995. Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering. *UCREL Technical papers*. Vol 5, University of Lancaster.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and coincidence. *Computational Linguistics*, 19(1):61-74.

Valentine Herman. 1974. What Governments Say and What Governments do: An Analysis of Post-War Queen's Speeches. *Parliamentary Affairs*, 28(1):22-31.

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

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

Dominique Labbé. 2007. Experiments on Authorship Attribution by Intertextual Distance in English. *Journal of Quantitative Linguistics*, 14(1):33-80.

Dominique Labbé and Denis Monière. 2008. *Les mots qui nous gouvernent. Le discours des premiers ministres québécois: 1960-2005.* Monière-Wollank, Montréal, QC.

Christopher D. Manning and Hinrich Schütze. 2000. *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, MA.

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

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

Kristina Toutanova, and Christopher Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagging. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, 63-70.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclid Dependency Network. In *Proceedings of HLT-NAACL 2003*, 252-259.