Published in Journal of Quantitative Linguistics 17, issue 2, 123-141, 2010
which should be used for any reference to this work

Lexical Analysis of US Political Speeches*

Jacques Savoy
Computer Science Department, University of Neuchatel, Switzerland

ABSTRACT

This article describes a US political corpus comprising 245 speeches given by senators John McCain and Barack Obama during the years 2007–2008. We present 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 analysed from both a statistical and dynamic perspective.

1. INTRODUCTION

The presidential election was the major political event in the United States in 2008. During this campaign the candidates (or their speech-writers) wrote various speeches that would hopefully convince undecided voters, to encourage their supporters and to make obvious that they were the best candidates 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, and 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

*Address correspondence to: Jacques Savoy, Computer Science Department, University of Neuchatel, rue Emile Argand 11, 2009 Neuchatel, Switzerland. Tel: +41 32 718 2700. Fax: +41 32 718 2701. E-mail: Jacques.Savoy@unine.ch

official version is usually the spoken one. But we can consider that the written version, usually available on each candidate's website, 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 article, 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 contains some conclusions.

2. RELATED WORK

In our analysis of political corpora and lexical analysis, we pay tribute to the work done by Labbé and 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 and 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 and Monière (2003), the reason for vocabulary increases cannot be attributed to a single and well-defined event, but may take place 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 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 US political corpus contains speeches we downloaded from the two candidates' official websites. For each speech, we added a few meta-tags to store document information (e.g. 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 upper-case letters by their corresponding lower-case, 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.

3.1 Overall statistics

Obama's speeches were downloaded from www.barackobama.com, beginning with the first on 10 February 2007 and ending with that on 30 October 2008 (Table 1 indicated the main dates of this election). In total our corpus contains 150 speeches (37 in 2007, 113 in 2008), for a total data size of 2.3 Mb (0.7 Mb for 2007, 1.6 Mb for 2008). For the Republican Party's speeches, we downloaded them from www.johnmccain.com beginning on 25 April 25 2007. This second subset contains 95 speeches (23 for 2007, 72 for 2008), for a total of 1.2 Mb (0.3 Mb for 2007, 0.9 Mb for 2008).

The data listed in Table 2 shows that McCain gave fewer speeches than Obama (95 vs. 150). 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

Table 1. Main events during the latest US presidential campaign.

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

compared over the last year (2007 mean: 3402; 2008 mean: 2607), and that they tended to have for year 2008 a mean length slightly larger than McCain's speeches (2174), who showed also a reduction during year 2008 (computation done with **R** [Crawley, 2007] and text processing with Perl [Nugues, 2006]).

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 7792 distinct word forms that McCain used in his speeches in 2008, 2958 (or

Table 2. Distribution of speeches by date and author.

	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	5	17
10/2008	15	24
Total	95	150

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

	McCain	Obama
Total of tokens	208,684	420,410
in 2007	54,319	125,857
in 2008	154,365	294,553
Tokens / speech	2200	2803
in 2007	2362	3402
in 2008	2174	2607
Number of forms	9014	9401
in 2007	5108	6547
Hapax in 2007	2171 (43%)	2476 (38%)
Frequency ≤ 4 in 2007	3699 (72%)	4411 (67%)
in 2008	7792	7663
Hapax in 2008	2958 (38%)	2573 (34%)
Frequency ≤ 4 in 2008	5146 (66%)	4617 (60%)

38%) word forms were used only once (a phenomenon known as *hapax*). Words used four times or less represent a rather large proportion, namely 66.1% of the total (or 5146 word 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 (95 vs. 150), his vocabulary tended to have a similar size (9014 vs. 9401).

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 (Francis & Kučera, 1982) (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

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

	Brown		US	
Rank	Lemma	Frequency	Lemma	Frequency
1	the	6.90%	the	4.69%
2	be	3.86%	be	3.81%
3	of	3.59%	and	3.78%
4	and	2.85%	to	3.30%
5	to	2.58%	of	2.61%
6	a	2.28%	that	2.17%
7	in	2.06%	a	1.95%
8	he	1.92%	in	1.88%
9	have	1.23%	we	1.85%
10	it	1.08%	I	1.50%
11	that	1.05%	have	1.36%
12	for	0.89%	not	1.19%
13	not	0.87%	for	1.18%
14	I	0.83%	our	1.10%
15	they	0.82%	it	1.01%
16	with	0.72%	will	0.98%
17	on	0.61%	this	0.85%
18	she	0.60%	you	0.68%
19	as	0.59%	do	0.65%
20	at	0.53%	on	0.62%

seem to play a really significant role and would thus not invalidate any comparisons.

From Table 4 it can be seen that "the" tends to occur more frequently in ordinary language (6.9%) than in political speeches (4.69%). What is more interesting is the conjunction "that" which ranks 6th in our US political speeches but only 11th in the Brown corpus. This tends to indicate 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 with 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 44th. The difference is even greater for the pronoun "she" (18th vs. 208th). Applying the Wilcoxon matched-pairs signed-ranks test (Conover, 1971) on data depicted in Table 4, we can verify whether both rankings reflect a similar words usage. In the current case, we must reject this hypothesis (significance level $\alpha = 0.05$, 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.

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- ($C = S \cup C$-), and contains n tokens ($n = a+b+c+d$).

Based on the MLE (maximum likelihood estimation) principle the values shown in a contingency table (see Table 5) could be used to

Table 5. Example of a contingency table.

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

estimate various probabilities. For example we might calculate the probability of the occurrence of the word ω in the entire corpus C as $\text{Prob}(\omega)=(a+b)/n$ or the probability of finding in C a word belonging to the set S as $\text{Prob(S)}=(a+c)/n$.

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 ($\text{Prob}(\omega)=(a+b)/n$). On the other hand, the probability of selecting a word in C belonging to the set S could be estimated by $\text{Prob(S)}=(a+c)/n$. Then if we consider selecting from C an occurrence of the word ω belonging to the set S, we could estimate this probability using $\text{Prob}(\omega \cap \text{S})=a/n$. However we could also assume that the joint event ($\omega \cap$ S) would be independent (by chance only) of both events (ω and S), which in turn would lead to another estimate, $\text{Prob}(\omega) \cdot \text{Prob(S)}$.

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

$$I(\omega;\text{S}) \;=\; \log_2\left[\frac{\text{Prob}(\omega \cap \text{S})}{\text{Prob}(\omega) \bullet \text{Prob(S)}}\right] = \log_2\left[\frac{a}{(a+b)} \bullet \frac{n}{(a+c)}\right] \qquad (1)$$

When the two estimates are close ($\text{I}(\omega, \text{S}) \approx 0$), 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 $\text{Prob}(\omega \cap \text{S}) > \text{Prob}(\omega) \cdot \text{Prob(S)}$, resulting in $\text{I}(\omega; \text{S}) > 0$. Finally, if $\text{Prob}(\omega \cap \text{S}) << \text{Prob}(\omega) \cdot \text{Prob(S)}$, this indicates that the two events are complementary and thus $\text{I}(\omega; \text{S}) < 0$. An example using this metric is given in Table 6 illustrating how the word ''IT'' is distributed in Obama's speeches in 2008 and in the rest of our US corpus.

From data depicted in Table 6, we could estimate directly the probability Prob($\omega \cap$ S) as 1/629,094 that corresponds to the numerator of Equation (1). On the other hand, we may estimate the probability of this joint event as the product of two independent events, namely Prob(ω) · Prob(S).

Using data in Table 6, we obtain (1/629,094) · (294,553/629,094) = $0.7 \cdot 10^{-6}$. The ratio of these two estimates is (1/629,094)/($0.7 \cdot 10^{-6}$) = 2.1357 from which we must take the logarithm according to Equation (1).

The resulting MI measure is I("IT"; Obama '08) = 1.09, 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 "IT" occurs just once in one Obama's speech in 2008 (as well as "Thanksgiving", "zionist", "Byron", or "astronaut"). According to our MI measure, this rare event returns a high MI value, tending to indicate a real

Table 6. Distribution of the word "IT" in Obama (2008) and US speeches.

	Obama '08	US-	
"IT"	1	0	1
Without "IT"	294,552	334,541	629,093
	294,553	334,541	629,094

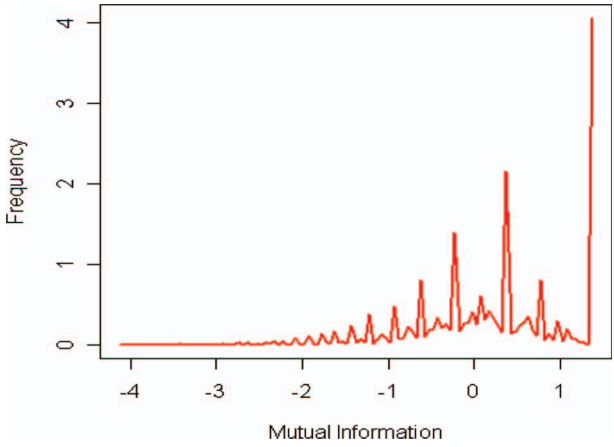

Fig. 1. Distribution of mutual information values: Obama 2008.

association between the word "IT" and Obama's vocabulary. Only one occurrence of this term can be found however and to ignore such particular cases, it is suggested that the additional constraint of $a \geq 5$ be imposed.

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 $n\,\text{Prob}(\omega) \cdot \text{Prob}(\text{S}) = n \cdot (a+b)/n \cdot (a+c)/n = (a+b) \cdot (a+c)/n$. 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 (e.g., a, b, etc.) and e_{ij} the expected frequency stored in cell ij.

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

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 7792 to 2646 (7792 – 5146) (or 34%) for McCain in 2008, and from 7663 to 3046 (or 39.8%) for Obama (see Table 2).

As shown in Table 7, the word "Bush" is distributed throughout McCain's speeches in 2008 and in the rest of our corpus. This word occurs 26 times in the subset and 398 times in the rest of the corpus. Under the assumption of independence, the expected frequency we should obtain is $(154{,}365 \cdot 424)/629{,}094 = 104$. For the three other cells

Table 7. Distribution of the word "Bush" in the McCain 2008 and US speeches.

	McCain '08	US-	
"Bush"	26	398	424
Without "Bush"	154,339	474,331	628,670
	154,365	474,729	629,094

of Table 7, we may repeat the computation of the expected frequencies and we would obtain the values depicted in Table 8.

The differences for the first cell (26 – 104) and the second (398 – 320) are rather large, indicating a divergence from the expected distribution. According to Equation (2), we must still raise these values to the power of 2 and divided them by the expected frequencies given in Table 8. For the first cell, we have $(26 - 104)^2/104 = 78^2/104 = 58.5$, and for the second we obtain $(398 - 320)^2/320 = 78^2/320 = 19$. The sum of these first two cells already gives 77.5. When considering the last two cells, the final sum value corresponding to χ^2 is quite high at 78.13.

Comparing this value with the limit value 6.63 ($\alpha = 0.01$, 1 dof, or 10.83 with $\alpha = 0.001$), we can reject the hypothesis that the word "Bush" 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. Senator McCain does not want to establish a clear link with the past president). This method owns the advantage of having a clear decision rule. We must, however, ignore a large set of words (around 64%, see Table 3) that occur fewer than five times in a subcorpus.

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 and Schütze (2000). This method could be appealing when faced with relatively low frequency values (e.g. less than five) 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).

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

Table 8. Expected frequencies of the word "Bush" in McCain 2008 and the rest of the corpus under the independence assumption.

	McCain '08	US-	
"Bush"	104	320	424
Without "Bush"	154,261	474,409	628,670
	154,365	474,729	629,094

We applied this measure in our corpus and Table 9 shows an example (the word ''the'' in Obama's 2008 speeches). The resulting G^2 value is 91.45, 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 (centred) and dividing it by its standard deviation (reduced). The resulting Z score value is also known as the standard score.

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

In Equation (4) we assume that the word ω follows 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 4).

From data depicted in Table 6, we can estimate p, the probability of occurrence of the word ''IT'' in the whole corpus as $1/629{,}094 = 0.0000016$, and n' is equal to 294,553. With these values, the corresponding Z score (''IT'') is

$$\text{Z score (''IT'')} = \left[\frac{1 - 294,553 \bullet (1/629,094)}{\sqrt{294,553 \bullet (1/629,094) \bullet (1 - (1/629,094))}}\right] = 0.777$$

In our opinion however the word distributions resembles the LNRE distributions (*Large Number of Rare Events* [Baayen, 2001]), and we

Table 9. Distribution of the word ''the'' in the Obama 2008 and US speeches.

	Obama '08	US-	
''the''	13,027	16,503	29,530
Without ''the''	281,526	318,038	599,564
	294,553	334,541	629,094

would therefore suggest smoothing the estimation of the underlying probability p as $(a+b+\lambda)/(n+\lambda \cdot |V|)$, where λ is a smoothing parameter (set to 0.5 in our case) and $|V|$ indicates vocabulary size (or 12,573 in the current case). This modification will slightly shift the probability density function's mass towards rare and unseen words (or words that do not yet occur) (Manning & Schütze, 2000).

In our previous example, the new estimate for p is $(1+0.5)/(629{,}094+0.5 \cdot 12{,}573)=0.0000024$, a value slightly larger than the previous one. The resulting Z score is also slightly different (0.365).

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, compared with 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 3.26%) or lower than -2 (5.13% for the current distribution). From applying this computation to the word ''Bush'' (Table 7), the resulting Z score is -7.6 clearly indicating that it is a word underused by Senator J. McCain. From Table 9, we observe the same conclusion; the determinant ''the'' has a Z score of -5.8 indicating an underused term in Obama's speeches during the year 2008.

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 ($-2 \leq$ Z score ≤ 2). It is, however, important to specify which corpus was used as reference. To do this 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 by Obama in 2007) or with only those speeches given by the same speaker (to verify how the author's vocabulary varies throughout the campaign).

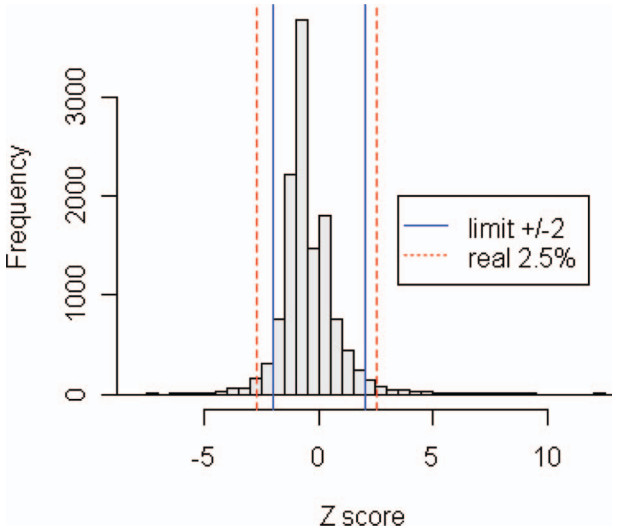

Fig. 2. Distribution of the Z score values: Obama 2008.

Table 10 lists the top overused and underused terms for both candidates, compared to the entire US political corpus. We examined all speeches (e.g. labelled ''McCain'') or only those speeches given in a specified year (e.g., only 2008 labelled ''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 the previous section). 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).

6. DYNAMIC ANALYSIS

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

Table 10. Terms overused and underused in speeches by Obama and McCain when compared with the entire corpus.

	Overused	Underused
McCain	government, Obama, honour, 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, honour, 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 11. Terms overused and underused in Obama's speeches when compared with the entire US corpus.

2008	Overused	Underused
January	deficit, Kennedy, Caroline, ...	government, energy, oil, McCain,
February	Orleans, NAFTA, FEMA, ...	oil, power, nuclear, security, ...
March	regulator, Wright, black, ...	energy, worker, oil, tax, ...
April	union, labour, worker, ...	war, nuclear, government, ...
May	Ryan, manufacturing, heroes, ...	nuclear, market, Iraq, ...
June	Israel, patriotism, cities, ...	politics, market, war, veteran, ...
July	Berlin, women, cyber, ...	politics, insurance, cost, Israel, ...
August	Joe Biden, McCain, oil, ...	war, reform, law, ...
September	financial, school, regulator, ...	war, Iraq, oil, ...
October	insurance, jobs, rescue, crisis, ...	war, nuclear, security, ...

entire US corpus and Table 12 lists all speeches delivered by the same speaker.

The contents of the two tables are fairly similar, revealing very little impact, regardless of whether we compared speeches with 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

Table 12. Terms overused and underused used by Obama in selected monthly speeches when compared with all his speeches.

2008	Overused	Underused
January	deficit, Kennedy, assumption, ...	McCain, million, energy, oil, ...
February	Orleans, NAFTA, FEMA, gulf, ...	world, oil, women, history, ...
March	regulatory, Wright, black, war, ...	you, energy, worker, tax, ...
April	labour, worker, union, trade, ...	war, school, education, ...
May	hemisphere, Cuba, Latin, freedom, ...	Iraq, kids, nuclear, market, ...
June	Israel, patriotism, Jewish, cities, ...	politics, war, veteran, people, ...
July	Berlin, women, cyber, Marshall, ...	politics, change, tell, story, ...
August	Joe Biden, oil, energy, renewable, ...	war, white, school, law, ...
September	financial, school, courses, McCain, ...	war, Iraq, oil, energy, women, ...
October	insurance, tax, rescue, jobs, ...	party, children, service, nuclear, ...

more traditional topics such as Pastor ''Wright'' in March, ''union'', ''labour'', ''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.

The Z score associated with the word ''Iraq'' changed for both candidates during the year 2008, as shown in Figure 3. 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 and October it tends to totally disappear from the campaign debate.

Clearly, as shown above in Figure 4, the term ''jobs'' is underused in 2007 by both candidates, while Obama reintroduced this question in the

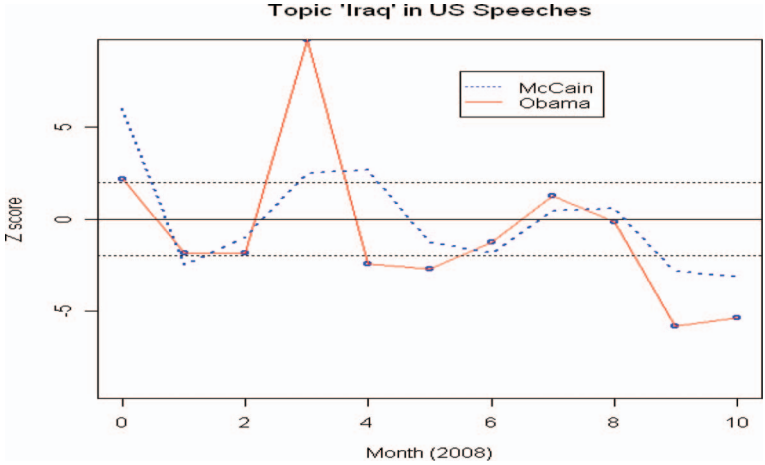

Fig. 3. Z score value for "Iraq" topic variations.

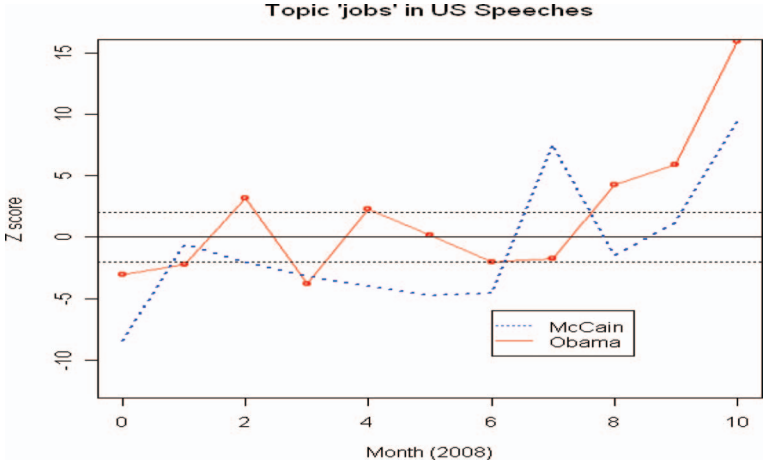

Fig. 4. Z score value variations for the topic "jobs".

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 and particularly in October 2008 both candidates tended to overuse this term. Comparing Figure 3 (topic "Iraq") and Figure 4 (topic "jobs"), we can see a clear shift from the foreign policy ("war in Iraq") to

more domestic problems related with the financial crisis and the ''jobs'' (see also Table 12).

8. CONCLUSION

In this paper we described the elaboration of a political corpus comprising 245 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 uncomfortable with a given topic.

REFERENCES

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

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

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

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

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

Daille, B. (1995). Combined approach for terminology extraction: Lexical statistics and linguistic filtering. *UCREL Technical Papers*. Vol 5. University of Lancaster.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

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

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

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

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

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

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

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

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

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

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

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