PART I

Computational methods for political text analysis

Introduction

Piek Vossen
VU University Amsterdam

Political language is one of the most challenging text types to analyze. In debates, politicians use language with great rhetorical skill. For computational analysis, political language is a genuine challenge. Natural language processing (NLP) by computers can be used to automatically analyze text and to derive implications from it, but all technology is limited to information that is structurally and statistically observable. There are implications that computers can draw from analyzing vast amounts of text, which people cannot do simply because they cannot store exact statistical data and cannot read so fast. Computers can, for example, detect trends in word usage across time and across different groups of (political) speakers that signify changes in political discourse. However, we also know that there are vast amounts of implications that humans can draw even from the smallest piece of text, but computers cannot do this. People can because they have an understanding of the complex social relations between the participants in the debate, have rich and complex background knowledge about the world we live in and are extremely experienced and sensitive to the use of language within such contexts. It is one of the exciting aspects of the field of text mining to see how far we can get in drawing implications from (political) text. Part I of this volume contains 6 articles that illustrate the possibilities and limitations of applying computational techniques to the analysis of political text. It is not a comprehensive overview, but the following chapters show some of the main issues that are currently under discussion.

In principle, the units and aspects of language can vary from individual words, to full text and collections of text, and from plain statistics to party issue positions and rhetoric. Table 1 gives an overview of the possible units of analysis (cf. first column) and their types of analysis (cf. first row).

Research and development is done on all these aspects but progress and the complexity of such efforts vary a lot. Statistics can be derived easily for any structure, but automatic analysis is more complex as we move on the scale from 'statistics' to 'world knowledge', and fewer systems are available that can do the job. In this book we see examples of NLP techniques centered to the left side of the table, whereas deeper qualitative approaches at the right side are necessarily restricted to a single or a small set of documents to be analyzed by humans (see Wesley, this volume).

Table 1. Units of text and types of text analysis in the field.

	Statistics	Structure	Meaning	Polarity	Position	Rhetoric	World knowledge
words	+	+	+	+	−	−	+
phrases	+	+	+	+	+	−	−
sentences	+	+	+/−	+	+	−	−
paragraphs	+	+	−	+		+	−
complete discourse or document level	+	+/−	−	+	+	+	−
text collections	+	+/−	−	+	−	−	−

In terms of methods, NLP has seen an impressive development over the past decades from rule-based and knowledge-rich systems, to machine-learning approaches, and most recently, to hybrid solutions. For various reasons, statistical and machine-learning methods are more accessible and widely used. One reason is their success over the traditional rule-based systems. Another reason is the light-weight and shallow processing of the text, which can be applied to large volumes. It has been shown that statistical and machine-learning NLP can, to a reasonable extent, predict party positions along political dimensions on the basis of language used by politicians. This book includes several examples of how this can be done.

In Chapters 2 (Collette and Pétry) and 5 (Hirst, Riabinin, Graham, Boizot-Roche and Morris), we see how the words of texts can be used to find party positions. In these approaches, the assumption is made that similar texts tend to use similar words and that there is no need to preserve the structure and complex composition of texts. The words that make up the text are the features that are 'machine learned' to make predictions. Collette and Pétry compare simple word-frequency methods of the widely-used programs Wordfish and Wordscores to English and French versions of Canadian party manifestos and evaluate them against expert surveys on party positions. They try to measure degrees of influence of different languages on party positioning and also consider the effects of word stemming (reducing the feature-space). The programs associate words and their frequencies with the party manifestos of one election and compare these with the frequencies in the manifestos of another election. Collette and Pétry show that this works equally well for English and French parliamentary debates in Canada, despite the different morphological properties of these languages. They also note that Wordscores outperforms Wordfish and that stemming words does not lead to significant effects.

Chapter 5 represents an interesting contrast with this paper. Hirst et al. also try to discover party positions in Canadian politics (liberal versus conservative) but use a Support Vector Machine (SVM) as a model and apply their analysis to

the English and French debates rather than to party manifestos. Hirst et al. come to the remarkable conclusion that their SVM classifier does not learn the language of the political position but merely the language of defense and attack. In their corpus, the liberals and conservatives swap places from opposition to government and vice versa. A classifier trained with the opposition language identifies the opposition regardless of the party's political status and the other way around for the language of the governing party. The results hold for both English and French. This chapter demonstrates that texts consist of many layers of information and it is dangerous to associate bags-of-words with any type of labels in classifiers, since we do not know what the classifier actually learns. In other words, we do not know which words belong to which layer of information. This is a genuine risk of any machine-learning approach, which can only be dealt with by rigorous testing and evaluation on many different data sets.

Another pair of chapters tries to extract positive/negative attitude or sentiment from heterogeneous types of text. In Chapter 3, Gryc and Moilanen compare different types of linguistic attitude indicators and social network data to derive sentiments centered around Barack Obama during the 2008 US elections. They apply their methods to 700 blog posts that are classified by crowd sourcing. Different feature sets are derived from the blog post (social network features, sentiment words, bags-of-words) and different classifiers are trained. Results are moderate, with the bags-of-words approach (using a large feature set) working best but social network analysis appearing to contribute additional evidence. Combining different classifiers through voting gives the best results, which is a well-known phenomenon in machine learning. Chapter 6 is closely related to this work: here, Grijzenhout, Jijkoun and Marx compare various techniques to determine subjectivity and polarity of paragraphs in Dutch parliamentary debates. They split the problem into determining: (1) the subjectivity of a paragraph and (2) the polarity of the subjectivity found. They compare different types of mathematical models for learning classifiers from training data and contrast the results with algorithms based on subjectivity lexicons. Both chapters show that state-of-the-art approaches to sentiment analysis (both machine-learning and lexicon-based) give reasonable results for political topics in various types of text, such as blogs and debates. Nevertheless, sentiment analysis for negative or positive attitude is rather one-dimensional compared to a complete analysis of the meaning and implication of political text.

Chapter 4 can be seen as an attempt to use similar techniques to perform a more complex task, namely to model the usage of the concept *outsiders* in the Swedish political debate. Dahlberg and Sahlgren use Random Indexing to measure concept-shifts. The basic idea is that statistics on the surrounding words tell you something about the meaning of a target word or concept. Whereas in the

previous techniques presented in this volume, sets of words are used to model larger text fragments associated with positions or sentiments, in this research the surrounding words (the word space) characterize the word *outsider* itself. In a first step, the language surrounding the concept *outsider* is learned from official documents of different parties. Next, the language used in a large collection of blogs is analyzed and compared with the *outsider* language of the parties. In this study, the analysis is complicated by lack of diachronic data. However, the *outsider* language of the parties tends to be similar and provides some evidence that the Conservative Moderate Party may have introduced connected concepts such as *unemployment* to the word space.

The chapters on computational methods show a strong tendency for statistical or quantitative approaches rather than for deeper qualitative approaches. In the light of Chapter 1, by Kleinnijenhuis and van Atteveldt, this kind of text analysis is relatively shallow. Even so, the conclusions and results of these shallow-quantitative techniques are still not without controversy. Results are moderate and probably not stable across different types of texts. Furthermore, Hirst et al. clearly show that we do not know what is learned since text, and definitely political text, comprises many different layers of information that may have been mixed up by the statistical analysis. However, it remains to be seen if deeper text analysis and more comprehensive approaches, as described by Kleinnijenhuis and van Atteveldt, can do a better job. For instance, the Net Method also leaves out many details and works by virtue of large volumes of news texts to wash out statistical value from noisy data. In order to become aware of the limitations and to evaluate the adequacy of computerized methods one should carefully consider and account for what is ignored.

To build a bridge between the (dis-)advantages of automated text analysis and methods for qualitative, interpretive analysis, Part I ends with a discussion of the fundamental differences between qualitative and quantitative methods for analyzing political texts that are the subject of Parts II and III (Wesley, this volume). Even though quantitative methods are often computerized and therefore assumed to be more objective, Wesley argues for qualitative approaches in which the interpretation of the researcher plays a role even when it leads to a bias in the interpretation of results. He claims that interpretive-subjective methods can result in productive insights as long as they are applied rigorously, following specified guidelines and choices are properly documented. Although not following the CDA paradigm, Wesley's chapter indicates a need for text analysis beyond quantitative premises. His approach builds a bridge to the chapters on qualitative discourse-analytic methods presented in Part II.

CHAPTER 2

Comparing the position of Canadian political parties using French and English manifestos as textual data

Benoît Collette and François Pétry
Université Laval, Department of Political Science

Recently, computer-assisted, quantitative methods have been developed to position political parties. These word-based textual analysis techniques rely exclusively on the relative frequency of words. As such they do not necessitate the knowledge of any particular language to extract policy positions from texts. However, different languages have different word distributions and other syntactic idiosyncrasies. These differences might provoke word-based textual analysis techniques to extract noticeably different positions from parallel texts that are similar in every aspect except language. How crippling is this potential disadvantage when comparing political texts written in different languages? It is this chapter's objective to determine the effect of language on the two word frequency methods Wordscores and Wordfish by comparing the policy positions of Canadian parties as extracted from their English and French party manifestos.

Word-based parallel content analysis

Over the past thirty years, the methods employed by researchers to locate political parties have evolved from hand-coded methods, such as the well-known Comparative Manifesto Project (CMP) (Budge et al. 2001; Klingemann et al. 2007), to expert surveys (Castles and Mair 1984; Huber and Inglehart 1995; Laver and Hunt 1992) to dictionary methods (Kleinnijenhuis and Pennings 1999; Laver and Garry 2000; Ray 2001). New computer-assisted, quantitative methods for extracting political party positions on the left-right axis or other policy dimensions from political texts have been a useful addition to the researcher's toolbox. They rely on objective textual data, they can be used to a nearly unlimited flow of data, and they make it possible to isolate policy preferences from behaviour (see Benoit and Laver 2007b; Laver and Garry 2000; Marks et al. 2007). One advantage

of word-based textual analysis techniques is that, since they rely exclusively on the relative frequency of words, they do not necessitate the knowledge of any particular language to extract policy positions from texts. However, different languages have different word distributions and other syntactic differences. These differences might provoke word-based textual analysis techniques to extract noticeably different positions from parallel texts that are similar in every aspect except language.

"Word-based" techniques, such as Wordscores (Laver et al. 2003) and Wordfish (Slapin and Proksch 2008), analyse the distribution of words in political texts to extract policy positions from them. We call these techniques "word-based" because the analytical units are the words in a text, not paragraphs, sentences, locutions or topics. This particularity has two main advantages. By chopping texts into words, word-based techniques gain the advantage of analytical simplicity, because words can be automatically identified and treated without human intervention.[1] Second, since words are treated like quantitative data, the knowledge of a language is no longer necessary to extract and then compare policy positions from texts written in different languages.

The disadvantage of word-based techniques is that they do not take into account the meaning or the grammatical structure of sentences and words that make them up. Focusing exclusively on the relative frequency of mention of words can lead to linguistic nonsense when the logic is pushed to its limits. As an extreme illustration, it is possible to extract a policy position from a random bunch of words that no human reader could make sense of, or from a freely reorganized text in, say, alphabetical order. It is impossible to measure the positive or negative direction of a policy preference in a text (Monroe et al. 2008). For example, the meaning of the sentence "We will raise taxes" is the exact opposite of "We will not raise taxes". But if we cut these sentences into separate words "we" (2 times) "will" (2) "not" (1) "raise" (2), and "taxes" (2), the difference between the two is "not", a relatively meaningless word which is likely to be overlooked. The difference in the meanings of the two sentences will be blurred as a result.

How crippling is this potential disadvantage when comparing political texts written in different languages? It is this chapter's objective to determine the effect of language by comparing the policy positions of parties in recent Canadian elections extracted from English and French party manifestos. We do this by checking whether two different methods for automated text analysis, Wordscores and Wordfish, extract the same policy positions on the left-right axis from parallel

1. In practice things are more complicated and experience showed us that, for example, hyphenated locutions can be sometimes treated as a single word or as separated words depending on the method used to produce a frequency matrix.

texts. Parallel texts are original documents written in different languages, not translations. They can be used to benchmark automatic translation quality (see Jian-Yun et al. 1999 for applications of parallel texts).

Do Wordscores and Wordfish extract the same policy positions on the left-right axis from parallel documents? In theory, there is no reason to believe that they would not. Parallel documents are rigorously the same. Their format is identical, they include the same topics and a bilingual reader will consider those documents to be almost exactly the same. Studies analyzing parallel textual data are interesting because they give an opportunity to test the validity and the reliability of word-based textual analysis methods in a more rigorous way than repeated studies focusing on a single language and/or a single party system. Wordscores has been tested with languages other than English, such as Dutch (Klemmensen et al. 2007), German (Hug and Schulz 2007; Magin et al. 2009; Bräuninger and Debus 2008), and French (Laver et al. 2006). But we could find only one study comparing Wordscores results using parallel texts as input. Debus (2009: 53–54) uses Wordscores to compare Flemish and French coalition agreements in Belgium and finds no significant difference between them for economic policy positions. In their analysis of European Parliament speeches Slapin and Proksch (2008) compare Wordfish results for speeches in English, French and German. They find remarkable similarities between languages (English and French especially):

> The comparison of the results across languages suggests that the position estimation technique is in fact highly robust to the choice of language (the correlation coefficient is 0.86 or higher). The highest correlation is between positions estimated from the English and French translations. These two languages are so similar to each other with regard to the information contained in words that they produce virtually identical position estimates. (Proksch and Slapin 2009: 13)

It should be noted that due to the large number of official languages spoken in the European Union, speeches in the European Parliament are delivered in one language and then translated. Slapin and Proksch relied on translations instead of original texts. It is unclear whether the relatively high level of similarity between texts in different languages was achieved because, or in spite, of the fact that they were translations. One way to clarify this is to extract party positions from parallel documents which original versions are written in more than one language.

Some specific features of languages can have a significant impact on the distribution of words in a text. French and English differ on many levels: syntax, grammar, and style (see Lederer 1994; Vinay and Darbelnet 2003 [1977]). An important syntactic difference between French and English is the use of articles. In French, they are more frequent than in English, because they are quasi-mandatory before

nouns. We can expect to find some common articles (de, du, la, les, etc.) to have a very high frequency, higher than the equivalent in English.

Looking at grammar, we find several significant differences. In French, grammatical gender results in a duplicate set of adjective, pronoun and article forms – one for masculine and one for feminine nouns. French texts tend to have relatively higher differentiations of those words when compared to English and that means expressing the same concepts with more words that are relatively less frequent than in English. While in English nouns may be singular or plural, in French adjectives can also be singular or plural, in addition of being masculine or feminine. So it is possible to have four word forms to express the same concept instead of one in English. As a simple illustration, take the adjective "pretty" in English. In French it can be declined in masculine singular form ("beau"), in feminine singular ("belle"), in masculine plural form ("beaux"), and feminine plural ("belles") depending on the object. Inevitably the English word "pretty" will be more frequent in a text than any one of the four French equivalent adjectives.

Another important difference between French and English is the verb tense system. In English, there are two forms in the present tense, one for the third person singular (he/she/it) and one for the rest, while most of the time in past tense you have only one form. The future, conditional and subjective tenses all have only one form, while French generally has five forms, one for each person plural (we/you/they) and two for the singular persons (either I/you or I/he-she-it), depending on the verb family. Moreover those forms are applied differently to future, conditional, subjunctive and past tenses. Therefore, the same verbs tend to be declined in more different words in French, modifying again the word distribution.

One consequence of these differences will appear in the number of words in a text. French texts have more words than English texts. Unsurprisingly, between 2000 and 2008 manifestos from major Canadian parties contain 11,717 different words in French, and 8,966 in English. The average frequency of words is higher in English than in French. On the surface, the average frequency of words is higher in French (19.1) than in English (18.7), but the standard deviation is much larger in French (218) than in English (176). Thirteen French words (du pour, un, en, d', l', le, a, des, la, et, les, de), many of them articles, appear more than 2,000 times, while in English only eight words appear more than 2,000 times (for, will, in, a, of, to, and, the). When we exclude those words the mean frequency becomes 12.8 in French and 14.3 in English. We also find more unique words (words used only once) in French. There are 4,625 unique words in French and 3,361 in English.

Methodology

In our study, we test Wordscores and Wordfish on Canadian parallel manifestos and compare the results on the left-right dimension with expert surveys to estimate the cross-language reliability of these two techniques. Canada, as an officially bilingual country, is a pertinent case for this test. At the federal level, manifestos from major parties are bilingual and both versions are considered official and public. With the exception of the Bloc Québécois, Canadian party manifestos can qualify as perfect parallel texts.[2] In addition, Canada has clearly identified polarized parties – the New Democratic Party (NDP) and the Conservative Party – on the left-right dimension (Irvine 1987; Johnston 2008). It is necessary to have a right and a left end point in the Canadian federal partisan spectrum in order to use Wordscores and Wordfish because both these techniques need to set values on reference texts.[3]

The analyses of both Wordscores and Wordfish are based on a frequency matrix. The frequency matrix lists the words used in all the documents and how frequent they occur in each document.[4] There is a debate over the merits of word stemming (reducing words to their root form) and removing stopwords (the, who, that, le, qui, quoi, etc.) with computer-assisted content analysis methods (see Lowe 2008). Stemming words and removing stopwords tends to reduce the size of texts and word diversity. That is supposed to lead to better results as meaningless words are removed and family of words reduced to stems. On the other hand, these operations artificially skew the word distribution in a text that could distort results. Although word stemming is a common practice with Wordfish, we will compare results with non-stemmed and stemmed texts. This comparison will be a pertinent

---

2. Canadian parties publish their platforms more or less simultaneously during election campaigns. One exception was the Liberals' "Green Shift" manifesto which was published well in advance of the 2008 election. The NDP, the Liberals and the Conservatives claim that the English and the French versions of their manifestos are original documents, not translations. This claim makes political sense, although it is impossible to verify. In any case, political parties know that any discrepancy between the English and the French versions of their manifesto would be advertized by the media and that this would undermine their credibility. The Bloc québécois does not produce sufficiently reliable English versions of its manifestos to qualify as parallel texts. The Bloc has been excluded from this analysis.

3. Contrary to what Slapin and Proksch (2008: 708) state, we need reference texts with Wordfish. Among the parameters, the user has to set omega scores for references texts representing both extremes of the dimension on which political texts will be located.

4. The frequency matrix from which data are taken has been produced with Will Lowe's jfreq program, available at http://www.williamlowe.net/software/ [Accessed November 3, 2010].

addition in the discussion. Jfreq has built-in stemming and stopword removal options that have been used to create the different matrixes needed.[5]

The corpus is composed of Canadian manifestos from the 2000, 2004, 2006, and 2008 federal elections. The relatively short period of time will minimize the problem of political vocabulary change, a concern for both Wordscores and Wordfish (Budge and Pennings 2007; Slapin and Proksch 2008). Three major parties are included in the analysis: the New Democratic Party (NDP), the Liberal Party and the Conservative Party.[6] Original manifestos have been tailored and standardized to facilitate content analysis. Messages from the leaders, financial annexes, tables, and quotations from other works (newspaper clips or OECD Reports for example) have been excluded from the original texts. For Wordfish, we followed Proksch and Slapin's (2008) recommendation and eliminated unique party words (words that are used by only one party) from the texts analyzed.

Canadian expert surveys

Expert surveys will serve as a benchmark to compare the face validity of the Wordscores and Wordfish estimates. Expert surveys are widely used to locate political parties in a policy space and often serve as a benchmark for textual analysis (Benoit and Laver 2007b; Keman 2007; Laver and Garry 2000; Marks et al. 2007; Ray 2007; Volkens 2007; Whitefield et al. 2007).[7] This approach was made popular thanks to a well publicized study by Castles and Mair (1984) who asked political scientists in 17 countries to locate political parties in their own country on an 11-point left-right scale. This was followed by a more ambitious study of left-right party position in 42 countries (Huber and Inglehart 1995). These two studies have inspired a large body of country-specific studies using expert surveys to locate political parties on a range of issues (for a sample see Laver 1998; Laver and Hunt 1992).

---

5. To remove French stopwords, we used the DotSEO list available at http://www.dot-seo.com/french-stop-words/ [Accessed November 3, 2010]. We had to tailor the list to fit this research. The modified list can be provided upon request by the authors.

6. The Reform Party/Canadian Alliance is not included because it no longer exists. It merged with the Progressive-Conservative Party in 2004 to make the Conservative Party. The Bloc Québécois is excluded because it does not publish English manifestos.

7. For reviews of the data on party placement and on self placement on policy spaces using the mass or expert survey method see Knutsen (1998); Budge (2000); Laver and Garry (2000); Mair (2001).

In constructing our own expert survey, we replicate the well-tested methodology already used by Laver and Hunt (1992), Laver (1998), Laver and Benoit (2005), Benoit and Laver (2007b). Our online survey questionnaire included 18 questions on a 0–10 scale. Respondents were asked to position the Bloc, Conservatives, Liberals, NDP, and Greens using the 11-point scale on the left-right axis and other policy dimensions. For all dimensions 0 represents the left position and 10, the right position. The expert survey was conducted immediately after the Canadian general election, held on October 14, 2008. The electronic questionnaire was sent to every political science professor in Canadian universities. 163 experts sent back the questionnaire out of 1,076 (15%).[8] Of the total, 127 questionnaires (78%) were filled in English and 27 (22%) in French.

Table 1. Canadian expert surveys in 2003 and 2008, 0–10 left-right scale.

Party	Mean both	Conf. interval	Mean English	Conf. interval	Mean French	Conf. interval	Laver/Benoit 2003 (s.e.)
Conservative	8.0	7.9–8.2	7.9	7.7–8.1	8.4	8.0–8.8	7.48 (0.11)
Liberal	5.3	5.0–5.6	5.2	5.0–5.4	5.5	5.2–5.8	6.1 (0.09)
NDP	2.9	2.7–3.1	2.9	2.7–3.1	3.0	2.5–3.6	2.7 (0.12)

Table 1 indicates how the experts who responded to our survey located the parties on the left-right policy dimension. Results from the Benoit and Laver survey (2006: 205), conducted in 2003 and included in the table, are quite similar.[9] The numbers in the first column are the mean scores for the entire sample of experts. But what is of direct interest here is the mean score for each subsample. The mean score for the position of the Conservative Party among English speaking respondents is 7.94 with a 95% confidence interval ranging from 7.7 to 8.1. If our sample of experts were a random sample (which it is not) we would say that 19 times out of 20 English speaking experts in the population located the Conservative Party between 7.7 and 8.1 points on the left-right axis in 2008. The mean score for the position of the Conservative Party among French speaking respondents is 8.39 with a 95% confidence interval ranging from 8.0 to 8.8. Note that the confidence interval for French speaking experts is much larger than for English speaking experts (this is of course due to the much smaller size of the French sample) and that the two intervals barely overlap each other. In other words, it is likely that French experts located the Conservative Party significantly more to the right than English experts in 2008. The mean scores for the position of the Liberal Party and

8. The data from the 2008 expert survey can be obtained upon request from the authors.

9. They report 104 respondents for a rate of 17% (Benoit and Laver 2006: 196).

the NDP are 5.2 for English experts and 5.5 for French experts, respectively. The small overlap between the two confidence intervals makes it once again likely that French speaking experts in the population locate the Liberals significantly more to the right than English speaking experts. The mean scores for the NDP differ very little in the sample and the large overlap of the confidence intervals indicate that the position of the NDP on a left-right axis by French speaking experts is indistinguishable from the position by English speaking experts. It should be noted that the scores for the Conservatives, the Liberals and the NDP are all statistically different from each other among both English and French speaking experts.

To summarize, both French and English speaking experts distinctly locate the Conservatives, the Liberals and the NDP. French speaking experts locate the Conservatives and the Liberals more to the right than English speaking experts, although the difference is not large for the Liberals. There is no difference in the way English and French experts locate the NDP.

Wordscores

The first computer-assisted method selected to locate the positions of political parties is the Wordscores computer program developed by Michael Laver and his co-researchers (Laver, Benoit and Garry 2003).[10] Unlike the CMP and dictionary methods that treat texts as discourse to be understood and interpreted for meaning either by a human coder or by a computer, Wordscores treats texts (more precisely the words contained in those texts) as data containing information about the position of the texts' authors on predefined policy dimensions. Starting from a set of "reference" texts whose policy positions are determined a priori, the technique extracts data from these reference texts in the form of word frequencies and uses this information to estimate the policy positions of "virgin" texts about which nothing is known.

Wordscores has the advantage of producing a distribution of scores around an estimated mean score. This makes it possible to calculate a standard error and hence to establish a confidence interval around the estimated mean score. Wordscores provides a statistical measure of how different two virgin texts are from one another in their vocabulary. Two texts are statistically different if their confidence intervals do not overlap. Of course, the scores are all the more valid if one has confidence in the choice of references texts and in the measure used to decide what their positions are on a given scale or cleavage. Laver and collaborators

10. For review of Wordscores, see Benoit and Laver (2007a, 2008); Klemmensen et al. (2007); Lowe (2008); Martin and Vanberg (2008); Slapin and Proksch (2008); and Volkens (2007).

make several important recommendations concerning the selection and scoring of references texts, one of them being that the virgin texts and the reference texts must share a similar frame of reference (Laver et al. 2003: 313–315). In our case, all documents are party manifestos that share the same structure. The recommendation is therefore fulfilled.

Here is a summary of how we used Wordscores to measure the positions of the Canadian parties on policy scales.[11] We first select the reference texts which are used to represent the extreme positions on the a priori defined policy scale. It is important that the reference texts are directly relevant to the virgin texts under analysis. We follow the common practice of using manifestos from the previous election as reference texts. The NDP manifestos are coded 0 (left), and the Conservative manifestos 10 (right) for the left-right dimension.

Each virgin text (that is each party platform at each election) is then coded by Wordscores which gives to each word in each virgin text a score between 0 and 10 according to the relative frequency of its appearance in the reference texts. For example, if the word "healthcare" appears one percent of the time in the NDP reference text, and 0.9 percent of the time in the Conservative reference text, "healthcare" obtains a score equal to (0.01*0) + (0.009*10) = 0.09. By dividing the sum of the scores associated with each word by the total number of words in a text, we obtain an average which corresponds to the total score of the text. From the wordscores in each reference text, we compute the textscores in each virgin text, and then transform the virgin textscores to their original metric to be able to locate the positions of each platform at each election in our pre-defined space.

Figure 1 and 2 show confidence intervals for English and French position estimates on the left-right dimension. The left-right ordering of the parties is consistent with the results of the expert survey. Note that the left-right ordering remains unchanged over time and that the shifts from one election to the other have the same direction in both the English and the French manifestos. As expected, the NDP is positioned on the left and the Conservative party on the right with the Liberal party somewhere between them. In Figure 1, it is impossible to statistically distinguish Liberal and Conservative parties in 2000 and 2004, as their confidence intervals overlap in these two elections. After 2004, the Conservative party moves to the right while the Liberal party moves to the left and gets closer to the NDP. The French Liberal platform which started in 2000 at the same point as the English Liberal platform does not move as far to the left in 2006 and in 2008. With stemmed

11. There is a debate over the proper score transformation method that should be used focusing mainly on comparability and scalability of transformed vs. raw scores (see Benoit and Laver 2008; Martin and Vanberg 2008). In this paper, we will keep the original method. This will make comparison easier, as most researchers who use Wordscores stick with the original method.

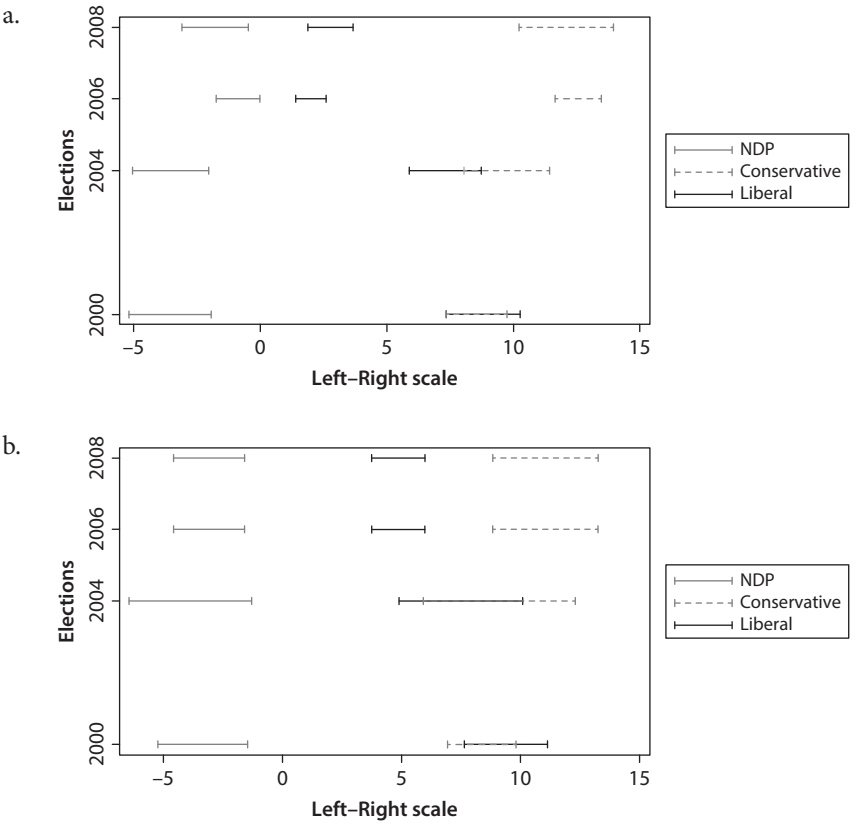


Figure 1. English (a) and French (b) left–right confidence intervals of Canadian manifestos with Wordscores.

documents, the leftward move of the Liberal party is so pronounced that the difference between the Liberals and the NDP tends to be blurred in 2006 and in 2008. The rapprochement between Liberal and NDP manifestos after 2004 in Figures 1 and 2 is consistent with the Canadian expert survey data of Table 1. According to Canadian experts, the average difference between the Liberals and the NDP decreased from 3.4 points in 2003 to 2.4 points in 2008 (and the average difference between the Liberals and the Conservatives increased from 1.4 points to 2.7 points).

English and French Wordscores estimates correlate at 97.1% with original documents (see the Appendix for details).[12] This figure goes down to 94.1% when stemmed documents are used. Overall, English documents, original or stemmed, have a higher correlation rate than French documents.

12. Debus (2009) does not indicate correlation levels for French and Flemish coalition agreements. Therefore we cannot establish a comparison.

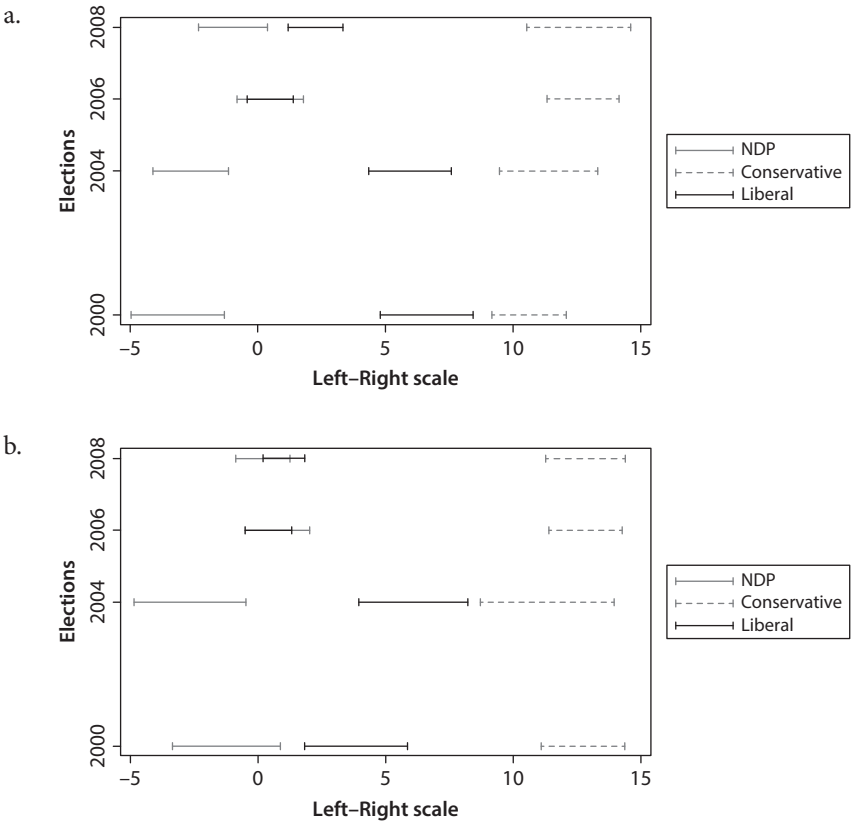


Figure 2. English (a) and French (b) left–right confidence intervals of stemmed Canadian manifestos with Wordscores.

Confidence intervals tend to be larger in French than in English: 3.8 against 2.5 with original documents and 3.3 against 3.0 with stemmed documents. We would expect longer documents to be more precise, as they contain more information. French manifestos are longer in general than English manifestos and they have more unique words and more total scored words. That means that although French manifestos contain more words, a significant number of them have median scores that tend to neutralize each other. For example, words like "the" or "and" are found multiple times in all texts. They will be given a median wordscore because of their ubiquity. Stemming has a positive impact in French by significantly reducing the number of ubiquitous and "meaningless" words and thus reducing the size of confidence intervals. The impact of stemming is more limited in English, as shown by the confidence intervals that are larger.

Wordfish

The Wordfish method takes a different approach. Borrowing from linguistics, it uses a naïve Bayes assumption to infer the process by which words are processed in a text.[13] A text is represented as a vector of word counts (occurrences) and individual words are assumed to be distributed at random. The probability that each word occurs in a text is independent of the position of other words in the text. While empirically false, naïve Bayes often performs well for classification (McCallum and Nigam 1998). Wordfish assumes the word frequencies are generated by a Poisson process. Slapin and Proksch (2008) chose this particular distribution because of its estimation simplicity. The single paramater, λ, is both the mean and the variance. The functional form of the model is as follows:

$$\lambda_{ijt} = \exp\left(\alpha_{it} + \psi_j + \beta_j * \omega_{it}\right)$$

Where λ_{ijt} is the count of word *j* in party *i*'s manifesto at time *t*, α is a set of party-election fixed effects, ψ is a set of word fixed effects, β is an estimate of a word specific weight capturing the importance of word *j* in discriminating between party position, and ω is the estimate of party *i*'s position in election year *t*. Each platform is treated as a separate party position and all positions are estimated simultaneously. That means there is no temporal constraint on the position of party *i*'s manifesto in election *t*. So if a party uses words in similar relative frequencies over time, its position will remain the same. Party movement is due to changes in word frequencies, not to change in word signification.

Wordfish gives two sets of results. The first is an estimation of the position of political parties on an axis that corresponds to the selected axis, with a confidence interval for every manifesto. The second is an estimate of the position of each word found in the selected texts. The document scores make it possible to position parties on policy dimensions and compare these positions with estimates of other methods. For reasons of comparability, we set the same values as Wordscores for reference texts in each dimension (0 for NDP 2004 and 10 for Conservative 2004). To calculate a 95% confidence interval, a parametric bootstrap, which is a re-sampling method that creates new datasets, is required.

Figure 3 compares the positions of English and French manifestos with non-stemmed documents. There is a similar pattern in both languages. It is statistically impossible to distinguish the positions of the parties in 2000, but at subsequent elections, they move apart from each other, with the Conservatives steering to the right and the Liberals to the left, with the NDP in between. This inversion of

13. See Lowe (2008) and Monroe et al. (2008) for a discussion over potential problems affecting Wordfish.

Liberal party and NDP – also present with stemmed documents in Figure 4 – is inconsistent with expert surveys and Wordscores results (both position the NDP on the left of the Liberals). Another striking feature of Figure 3 is the difference in the size of confidence intervals between English and French manifestos. Since French documents are longer, we would expect them to contain more information that logically lead to smaller confidence intervals, but results show the opposite (7.0 against 3.6 in average). The difference is especially significant for the Liberal party (0.46 in English against 12.6 in French), which is even more surprising, considering the fact that Liberal manifestos are longer than other parties' in both languages. The last significant difference lies in the relative position of parties along the scale. English manifestos tend to score between 0 and 10, close to the original scale. French manifestos score over a range that reaches much further to the left. In left-right terms, French manifestos are farther apart from each other.

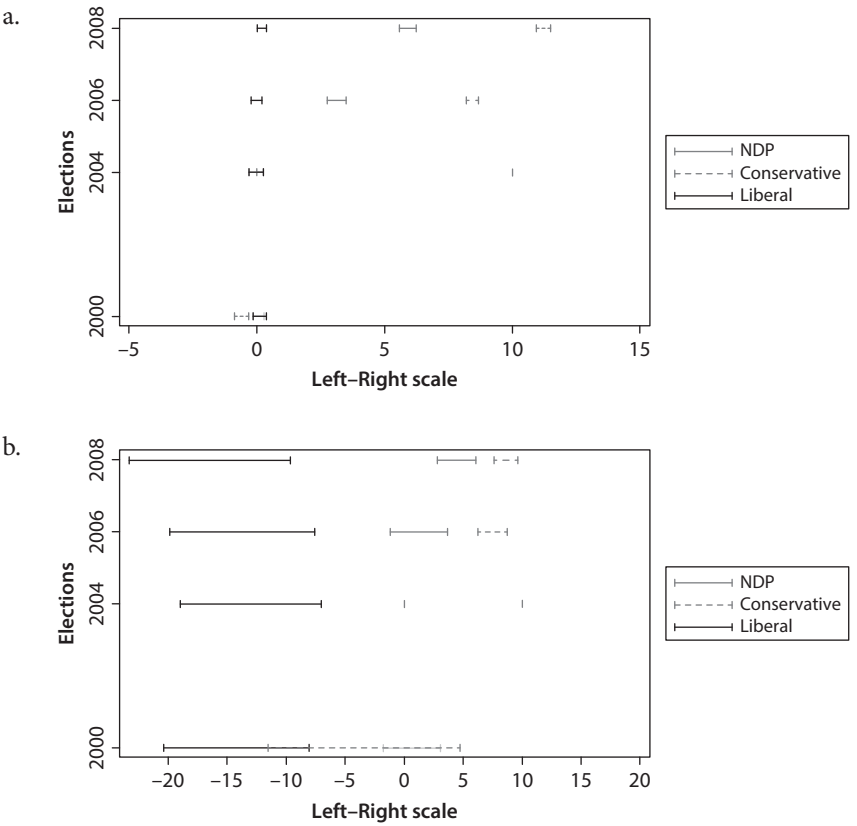


Figure 3. English (a) and French (b) left–right scores of Canadian manifestos with Wordfish.

Figure 4 compares the positions of English and French manifestos with stemmed words. Here the positioning in English and French is more similar. The correlation between languages goes from 70.2 with original words to 98.4 with stemmed words (see Appendix). This is comparable to results by Proksch and Slapin (2009: 19), who reported a 86% correlation between English and French stemmed speeches. Stemming also has a positive impact on confidence interval size (3.6 in French and 1.5 in English) with the gap diminishing between the two languages, from 3.4 to 2.1. While scales would vary considerably in terms of size between French and English with original documents, they have now similar sizes with stemmed documents, ranging more or less from −10 to 10 in English from −15 to 10 in French.

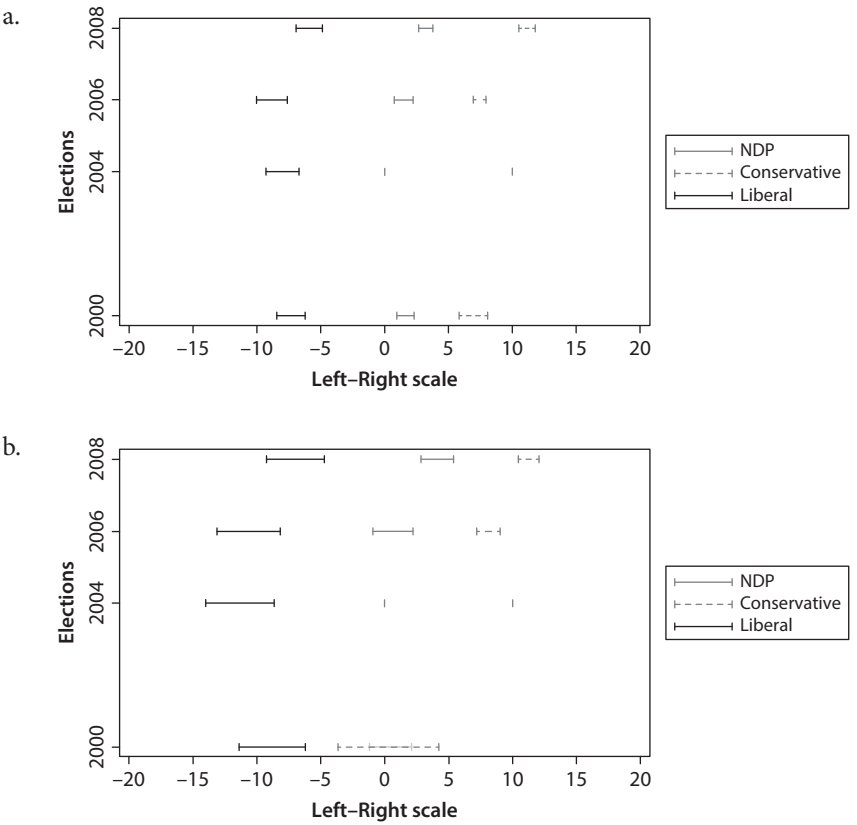


Figure 4. English (a) and French (b) left–right scores of stemmed Canadian manifestos with Wordfish.

Overall, stemming seems to have a positive impact, as confidence intervals shrink and correlation is higher between languages. The impact is more significant in French as we have seen. For this kind of comparison between languages, we would follow the recommendation to proceed with stemmed documents.

Conclusion

Do grammatical differences between languages threaten the validity and the reliability of left-right party position results generated by word-based textual analysis techniques such as Wordscores and Wordfish? Here is a summary of what we found. Wordfish incorrectly positions the Canadian parties. The positioning of the Liberal party on the far left of the political spectrum and the NDP close to the center in both languages is in contradiction with expert surveys and Wordscores estimates. Furthermore, it is also in contradiction with the perception of the Canadian electorate and CMP results (Budge et al. 2001; Klingemannn et al. 2007). Stemming the documents improves Wordfish results somewhat. The estimates of the English and French stemmed documents are more highly correlated and have smaller confidence intervals than the non-stemmed documents. However stemming and stopword removal do not solve the incorrect positioning of the NDP and the Liberal party in either language. In this case, we cannot judge whether grammatical differences threaten the validity and reliability of results with non-stemmed documents. In both languages they could not be considered as valid, because the absolute position of parties is significantly different. With stemmed documents, scales are more comparable, but estimates are still not valid. Of course, increasing the number of cases and looking at more specific policy dimensions could lead to different results. Maybe we are not positioning the parties on the left-right dimension after all. That is the problem: we are not sure on what dimension we are positioning parties when using Wordfish on whole manifestos.

With Wordscores, results are more consistent with expert surveys' estimates than with Wordfish. Interestingly we see the Liberals steering to the left over time, especially with stemmed documents. Stemming does not greatly alter Wordscores results. Apparently Wordscores can deal easily with "meaningless" words, as their ubiquity tends to give them scores close to the median that do not have a significant effect on text score. Therefore, stemming does not influence the results significantly. The grammatical differences of the two languages do not seem to threaten the validity and reliability of left-right party position estimates. The correlation between the estimates of the English and French manifestos is an impressive 0.97 for non-stemmed documents and 0.94 for stemmed documents. In both

English and French the positions of the manifestos shift in the same direction from one election to the other. Compared to Wordfish confidence intervals overlap more, but we maintain that Wordscores results are more valid and reliable than Wordfish estimates.

In this chapter, we have limited our scope to two European languages that are relatively close to each other. Including non-Indo-European languages in comparative research would be an interesting addition and a test for the universality of computer-assisted content analysis. The inclusion of languages that have a different morphological structure, such as isolating languages (e.g. Chinese), with low morpheme-per-word ratio, would be a good test of the ability of computer-assisted content analysis to discriminate text positions based solely on word distribution. Research in comparative linguistic is progressing fast (for example, see Yang and Li 2003) and its applications could be useful for political textual analysis.

References

Benoit, K. and M. Laver. 2006. *Party Politics in Modern Democracies*. New York: Routledge.

Benoit, K. and M. Laver. 2007a. Benchmarks for text analysis: A response to Budge and Pennings. *Electoral Studies* 26(1), pp. 130-135. DOI: 10.1016/j.electstud.2006.04.001

Benoit, K. and M. Laver. 2007b. Estimating party policy positions: Comparing expert surveys and hand-coded content analysis. *Electoral Studies* 26(1), pp. 90–107. DOI: 10.1016/j.electstud.2006.04.008

Benoit, K. and M. Laver. 2008. Compared to what? A comment on "A robust transformation procedure for interpreting political text" by Martin and Vanberg. *Political Analysis* 16(1), pp. 101–111. DOI: 10.1093/pan/mpm020

Bräuninger, T. and M. Debus. 2008. Der Einfluss von Koalitionsaussagen, programmatischen Standpunkten und der Bundespolitik auf die Regierungsbildung in den deutschen Ländern. *Politische Vierteljahresschrift* 49(2), pp. 309–338. DOI: 10.1007/s11615-008-0101-6

Budge, I., H.-D. Klingemann, A. Volkens, J. Bara and E. Tanenbaum (eds). 2001. *Mapping Policy Preferences I. Estimates for Parties, Electors, and Governments 1945–1998*. Oxford: Oxford University Press.

Budge, I. 2000. Expert judgments of party policy positions: Uses and limitations in political research *European Journal of Political Research* 37(1), pp. 103–113. DOI: 10.1111/1475-6765.00506

Budge, I. and P. Pennings. 2007. Do they work? Validating computerized word frequency estimates against policy series. *Electoral Studies* 26(1), pp. 121–129. DOI: 10.1016/j.electstud.2006.04.002

Castles, F.G. and P. Mair. 1984. Left-right political scales: Some 'expert' judgments. *European Journal of Political Research* 12(1), pp. 73–88. DOI: 10.1111/j.1475-6765.1984.tb00080.x

Debus, M. 2009. Pre-electoral commitments and government formation. *Public Choice* 138(1), pp. 45–64. DOI: 10.1007/s11127-008-9338-2

Huber, J.D. and R. Inglehart. 1995. Expert interpretations of party space and party locations in 42 societies. *Party Politics* 1(1), pp. 73–111. DOI: 10.1177/1354068895001001004

Hug, S. and T. Schulz. 2007. Left-right positions of political parties in Switzerland. *Party Politics* 13(3), pp. 305-330. DOI: 10.1177/1354068807075938

Irvine, W. 1987. Canada 1945–1980: Party platforms and campaign strategies. In I. Budge, D. Robertson and D. Hearl (eds), *Ideology, Strategy, and Party Change: Spatial Analysis of Post-War Elections Programmes in Nineteen Democracies*. Cambridge: Cambridge University Press, pp. 73–94. DOI: 10.1017/CBO9780511558771.005

Jian-Yun, N., M. Simard, P. Isabelle and R. Durand. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. Berkeley, California, United States: ACM.

Johnston, R. 2008. Polarized pluralism in the Canadian party system: Presidential address to the Canadian Political Science Association, June 5, 2008. *Canadian Journal of Political Science/ Revue canadienne de science politique* 41(4), pp. 815–834.

Keman, H. 2007. Experts and manifestos: different sources – Same results for comparative research? *Electoral Studies* 26(1), pp. 76–89.

Kleinnijenhuis, J., and P. Pennings. 1999. A probabilistic keyword approach to textual analysis. In *Mannheim Joint Sessions of the ECPR*. Mannheim.

Klemmensen, R., S. B. Hobolt and M. E. Hansen. 2007. Estimating policy positions using political texts: An evaluation of the wordscores approach. *Electoral Studies* 26(4), pp. 746–755. DOI: 10.1016/j.electstud.2006.04.004

Klingemann, H.-D., A. Volkens, J. Bara, I. Budge and M. D. McDonald. 2007. *Mapping Policy Preferences II: Estimates for Parties*, Electors and Governments in Central and Eastern Europe: Oxford University Press.

Knutsen, O. 1998. The strength of the partisan component of left-right identity: A comparative longitudinal study of left-right party polarization in eight west European countries. *Party Politics* 4(1), pp. 5–31. DOI: 10.1177/1354068898004001001

Laver, M. 1998. Party policy in Britain 1997: Results from an expert survey. *Political Studies* 46(2), pp. 336–347. DOI: 10.1111/1467-9248.00144

Laver, M. and K. Benoit. 2005. Estimating party policy positions: Japan in comparative context. *Japanese Journal of Political Science* 6(2), pp. 187–209. DOI: 10.1017/S1468109905001830

Laver, M., K. Benoit and J. Garry. 2003. Extracting policy positions from political texts using words as data. *The American Political Science Review* 97(2), pp. 311–331. DOI: 10.1017/S0003055403000698

Laver, M., K. Benoit and N. Sauger. 2006. Policy competition in the 2002 French legislative and presidential elections. *European Journal of Political Research* 45(4), pp. 667–697. DOI: 10.1111/j.1475-6765.2006.00313.x

Laver, M. and J. Garry. 2000. Estimating policy positions from political texts. *American Journal of Political Science* 44(3), pp. 619–634. DOI: 10.2307/2669268

Laver, M. and W. B. Hunt. 1992. *Policy and Party Competition*. New York: Routledge.

Lederer, M. 1994. *La traduction aujourd'hui: le modèle interprétatif*. Vanves: Hachette F. L. E.

Lowe, W. 2008. Understanding wordscores. *Political Analysis* 16(4), pp. 356–371. DOI: 10.1093/pan/mpn004

Magin, R., M. Freitag and A. Vatter. 2009. Cleavage structures and voter alignments within nations. *Zeitschrift für Vergleichende Politikwissenschaft* 3(2), pp. 231–256. DOI: 10.1007/s12286-009-0062-1

Mair, P. 2001. Searching for the positions of political actors: A review of approaches and a critical evaluation of expert surveys. In M. Laver (ed.), *Estimating the Policy Positions of Political Actors*. New York: Routledge, pp. 10–29.

Marks, G., L. Hooghe, M. R. Steenbergen and R. Bakker. 2007. Crossvalidating data on party positioning on European integration. *Electoral Studies* 26(1), pp. 23–38. DOI: 10.1016/j.electstud.2006.03.007

Martin, L. W. and G. Vanberg. 2008. A robust transformation procedure for interpreting political text. *Political Analysis* 16(1), pp. 93–100. DOI: 10.1093/pan/mpm010

McCallum, A. and K. Nigam. 1998. A comparison of event models for naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*.

Monroe, B. L., M. P. Colaresi and K. M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16(4), pp. 372–403. DOI: 10.1093/pan/mpn018

Proksch, S.-O. and J. B. Slapin. 2009. Position taking in European parliament speeches. *British Journal of Political Science* 40, pp. 587–611. DOI: 10.1017/50007133409990299

Ray, L. 2001. A natural sentence approach to the computer coding of party manifestos. In M. Laver (ed.), *Estimating the Positions of Political Actors*. New York: Routledge, pp. 149–161.

Ray, L. 2007. Validity of measured party positions on European integration: assumptions, approaches, and a comparison of alternative measures. *Electoral Studies* 26(1), pp. 11–22. DOI: 10.1016/j.electstud.2006.03.008

Slapin, J. B. and S.-O. Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science* 52(3), pp. 705–722. DOI: 10.1111/j.1540-5907.2008.00338.x

Vinay, J.-P. and J. Darbelnet. 2003 [1977]. *Stylistique comparée du français et de l'anglais : Méthode de traduction*. Laval: Groupe Beauchemin.

Volkens, A. 2007. Strengths and weaknesses of approaches to measuring policy positions of parties. *Electoral Studies* 26(1), pp. 108–120. DOI: 10.1016/j.electstud.2006.04.003

Whitefield, S., M. A. Vachudova, M. R. Steenbergen, R. Rohrschneider, G. Marks, M., P. Loveless, and L. Hooghe. 2007. Do expert surveys produce consistent estimates of party stances on European integration? Comparing expert surveys in the difficult case of Central and Eastern Europe. *Electoral Studies* 26(1), pp. 50–61. DOI: 10.1016/j.electstud.2006.04.006

Yang, Ch. C. and K. Wing Li. 2003. Automatic construction of English/Chinese parallel corpora. *Journal of the American Society for Information Science and Technology* 54(8), pp. 730–742. DOI: 10.1002/asi.10261

Appendix

Document scores correlations

Wordscores	English	French	English (stemmed)	French (stemmed)
English	1.000			
French	0.9710	1.000		
English (stemmed)	0.9926	0.9509	1.000	
French (stemmed)	0.9374	0.8983	0.9411	1.000

Wordfish	English	French	English (stemmed)	French (stemmed)
English	1.000			
French	0.7020	1.000		
English (stemmed)	0.9390	0.8009	1.000	
French (stemmed)	0.9242	0.8154	0.9843	1.000

CHAPTER 3

Leveraging textual sentiment analysis with social network modeling

Sentiment analysis of political blogs in the 2008 U.S. presidential election

Wojciech Gryc and Karo Moilanen
Oxford Internet Institute, University of Oxford /
Oxford University Computing Laboratory

Automatic computational analysis of political texts poses major challenges for state-of-the-art Sentiment Analysis and Natural Language Processing tools. In this initial study, we investigate the feasibility of combining purely linguistic indicators of political sentiment with non-linguistic evidence gained from concomitant social network analysis. The analysis draws on a corpus of 2.8 million political blog posts by 16,741 bloggers. We focus on modeling blogosphere sentiment centered around Barack Obama during the 2008 U.S. presidential election, and describe a series of initial sentiment classification experiments on a data set of 700 crowd-sourced posts labeled for attitude with respect to Obama. Our approach employs a hybrid machine-learning and logic-based framework which operates along three distinct levels of analysis encompassing standard shallow document classification, deep linguistic multi-entity sentiment analysis and scoring and social network modeling. The initial results highlight the inherent complexity of the classification task and point towards the positive effects of learning features that exploit entity-level sentiment and social-network structure.

1. Introduction

Political blogs constitute a fascinating genre. They are typically charged with extremely high amounts of personal opinions, sentiments, stances, feelings and emotions towards and about a multitude of individuals, organisations, issues and events that have some political relevance, and represent a very large number of people. Due to the sheer size of such a complex distributed, dynamic and vibrant information space, the only viable way of monitoring, analysing and predicting

information in the political blogosphere is to develop a large-scale computational multi-paradigm and multi-modal framework that can deal with the content as well as the social aspects of the blogosphere.

With this lofty goal in mind, we investigate in this study the feasibility of combining purely textual indicators of political sentiment with non-textual information gained from concomitant social network analysis. We focus on political blog analysis and base our research on a large corpus of posts by 16,741 bloggers, crawled daily between April 2008 and May 2009. Rich political blog data allows us to explore a number of major themes in political sentiment analysis, social network analysis, text mining, and, most importantly, how these areas interact. For example, by exploring the content of each post and how it fits into the larger social network in which it resides one can see that discussions tend to cluster differently depending on individual politicians: people discussing Barack Obama tend to be more dispersed than those discussing John McCain, for example. An interesting and highly relevant research question is whether such network-based information can be used to improve the accuracy of automatic political-sentiment classification.

In this initial study, we approach automatic sentiment analysis of political blogs using a hybrid machine-learning and logic-based classification framework which operates along three distinct complementary levels of analysis, each capable of offering a unique representation of each blog post:

- *Shallow document classification* (§4.1): as a first strong baseline, centered around holistic document-level textual evidence alone, each blog post is represented using standard *n*-gram features.
- *Deep entity-level sentiment scoring* (§4.2): in a second approach centered around much more focused lower-level sentiment evidence, each blog post is represented as a set of detailed sentiment scores assigned to all individual entities mentioned in the post (e.g., politicians, places, organisations, and abstract issues).
- *Social network modeling* (§4.3): in a third, much wider meta-approach centered around the blog posts' social context, each blog post's position in the social network structure across the whole blog-post space is used as classification evidence.

Classification task

Although our framework is not restricted to any particular topic or entity (be they concrete or abstract), we confine[1] ourselves in this initial study to only one specific entity in order to gain a better understanding of the problem. In light of the

1 We aim to expand this in future to further politicians and topics such as *McCain* and *Iraq*.

importance of social media to Barack Obama's political campaign in the 2008 U.S. presidential election, we decided to focus our analysis on the sentiment expressed in the blogosphere towards and about Barack Obama within the time period that our blog data represents. The classification task that we attempt in this initial study can be summarised specifically in the following question:

> *Given a political blog post, does it, as a whole, express positive, neutral, or negative sentiment towards or about the target entity (Barack Obama)?*

The classification task can be characterised accordingly as *entity-centric document-level sentiment classification* because the document-level sentiment polarity label of a given post reflects the overall sentiment towards or about a single target entity in that post, not the overall sentiment of the post as a whole. Note that this type of sentiment classification is different from the 'traditional' document-level classification paradigm because a post that is negative *overall* (i.e., as a document) can (and is likely to) be concurrently negative, positive, or neutral towards or about many individual entities. We approach the problem as if we were trying to label the entire data set after it was collected, rather than in real time. We are more interested in being able to label a majority of the 2.8 million posts we have collected with some level of confidence rather than having to look into the future when labeling specific posts.

General challenges

Automatic computational analysis and categorisation of political texts is on the whole a seriously challenging task, as has been observed in the area (e.g., Durant and Smith 2007; Malouf and Mullen 2008; Mullen and Malouf 2006; Thomas et al. 2006; Yu et al. 2008). When the analytical scope is extended to include further non-factual aspects of meaning pertaining to subjectivity, sentiment, opinions, affect and emotions, the analytical and computational challenges become even more pronounced. This is especially the case with politically charged content in blogs because, as a genre, political blogs represent noisy, in-depth, collaborative, and dynamic discussions and debates by multiple contributors across a wealth of topics, issues and entities – only some of which constitute core content while some others are mere digressing or tangential content. Blog posts further link to each other via highly complex interrelated direct/explicit and indirect/implicit structural, semantic, rhetorical, and temporal chains. It can be argued that no such chain can be explained fully out of context, although blog posts are likely to carry at least some clues that an algorithm can exploit. In addition to their distributed nature, blog posts can also include other forms of multimedia such as embedded videos or images which existing Natural Language Processing (NLP) algorithms cannot easily align with the text content.

From the viewpoint of textual sentiment-analysis algorithms, any classification evidence that may be gleaned from blog posts is inherently noisy because bloggers' real sentiments and opinions are often obfuscated by complex rhetoric, irony, sarcasm, comparisons, speculation and other paralinguistic devices that prototypically characterise political discussions. In addition, domain-specific terms, word senses, and vocabularies and informal/non-standard registers also feature frequently. As is the case with Web content in general, political blogs also come with many purely textual hurdles that have to be faced by NLP tools such as complex or incomplete grammatical structures, broken sentence boundaries, quotes, junk characters and spelling anomalies.

Even if such structural and textual problems were solved fully, the very task of automatically detecting bloggers' political sentiments, opinions, and orientation pertaining to highly polarised political issues would still remain formidable. This stems from the fact that current computational tools – be they linguistic or non-linguistic – struggle to map raw surface clues onto deeper semantic representations which ultimately require, *for each blogger and for each issue, bill, or event* under consideration, the ability (i) to detect the blogger's political party affiliation, political viewpoints, professional background, motivation, general knowledge and, indeed, affective states; (ii) to measure the blogger's political extremeness or distance from a centrist position; (iii) to measure the blogger's confidence and agreeability/argumentativeness in the discussions; (iv) to measure how politically important something is to the blogger; (v) to understand how meaning was constructed collaboratively by the bloggers; (vi) to understand why certain topics are (not) discussed; and (vii) to detect sincere opinions vs. deliberate flaming. Moreover, not only are many political opinions latent behind expressions which (from an algorithm's point of view) present themselves as purely neutral but some explicit political opinion expressions may even be inversely related to the blogger's actual political orientation.

2. Data

2.1 Election data

The present initial study, which is part of a wider research project, focuses on data provided by IBM's *Predictive Modeling Group*. In total, the data set consists of 2,782,356 posts written by 16,741 politically-oriented bloggers collected between April 22nd 2008 and May 1st 2009 (many of which focus on the U.S. Presidential election in 2008). The selection of blogs was based on the tags associated with

them on the *Technorati*[2] blog indexing service. Such a blog data set provides an interesting opportunity for researchers because it contains text-based discussions on politics as well as date stamps and hyperlinks between blog posts. The hyperlink feature can be treated as a citation network which shows the various ways in which individual bloggers are and become aware of each other, and how information flows within the political blogosphere.

Posts were found through blogs self-reporting new content through the RSS (*Real Simple Syndication*) standard, which includes hyperlinks to the content posted on the blogs themselves. Once the blogs were crawled, the content of each post was filtered to discard text and hyperlinks from advertisements, side bar content (e.g., blog rolls) and other portions of the web pages that include unwanted superfluous content.

The role of content filtering is extremely important, mainly for three reasons. Firstly, blog titles and content from sidebars can easily bias any statistical models trained on noisy data. Since we are only interested in the main content of every blog post, it is important to be able to avoid such bias by discarding as many sidebar links as possible. Secondly, superfluous content can, even in structural terms, cause potentially devastating complications for NLP tools which can manifest themselves as incorrect sentence breaking, part-of-speech tags, phrase chunking, and parsing – all of which will deteriorate the performance of sentiment classifiers that use NLP components. Lastly, blogs often acted in an automated fashion. For example, blog A citing blog B can automatically cause blog A to leave a link as a comment on blog B's post and hence result in additional links. Blog rolls, while providing useful social-network information, may cause the network to become artificially dense with hyperlinks that do not pertain to the blog posts themselves. It is important to note, however, that while our filtering strategy is effective enough for practical purposes, it is based on heuristics and is by no means perfect.

2.2 Sentiment annotations

While the aforementioned data collection process allows us to analyse the content of the posts and to build and model social networks of bloggers and their posts, the crawling process did not as such deal with the sentiment expressed in the posts. In order to gain access to the sentiment properties of the posts, we made use of Amazon's *Mechanical Turk*[3] (MT) service which involves *Requesters* (us) posting

2. http://technorati.com/, last accessed March 6, 2014.

3. https://www.mturk.com/mturk/welcome, last accessed March 6, 2014.

batches of small tasks known as HITs (*Human Intelligence Tasks*) which are completed by *Workers* (*Turkers*) for a small monetary reward ($0.03 per blog post).

For the present initial study, we selected a random sample of 700 blog posts discussing Barack Obama and asked the Turkers to label them as (i) POSITIVE, NEGATIVE, NEUTRAL towards or about Obama, or (ii) NOT APPLICABLE (with respect to Obama). In order for us to monitor and ensure the post-level consistency of the Turkers' sentiment ratings, we required each post to be labelled by three Turkers. In total, 86 unique Turkers took part in the task. All posts labeled as NOT APPLICABLE were discarded because such cases typically indicate that a given blog post to be labeled had been taken offline or contained irrelevant non-textual content (e.g., a video or an image). From the resulting raw sentiment ratings, two labeled subsets were generated as follows:

- *Lenient majority vote*: The first subset contains all posts that received a majority vote whereby either 2/3 or 3/3 Turkers had to agree on the label of each post. This resulted in 454 posts from which a final[4] labelled corpus of 439 posts was obtained. All results reported in the present initial study are from this subset.
- *Strict agreement:* The second subset contains all posts that received unanimous 3/3 votes. Since that criterion resulted in only 124 posts, this subset is not included in the present study.

We are in the process of increasing the amount of sentiment annotations.

- *Human performance*. Even though it is complicated to estimate the inter-annotator agreement rates between 86 annotators (each of whom provided different amounts of annotations), some tentative observations can be made regarding the expected human agreement and performance ceiling in the sentiment classification task that our classifiers aim at solving. Table 1 shows the distribution of votes per sentiment polarity, taking 2/3 and 3/3 agreement ratings and excluding NOT APPLICABLE cases. It interestingly reveals that only 126 (27.75%) display FULL AGREEMENT, 145 (31.94%) display disagreements that involve neutral polarity (NEUTRAL DISAGREEMENT) and 183 (40.31%) display FATAL DISAGREEMENT in the form of opposing non-neutral polarities. Both the noticeably low amount of FULL AGREEMENT and the relatively high amount of FATAL DISAGREEMENT suggest that, perhaps not surprisingly, the 3-way classification task is highly subjective even for humans. It is against the strict ceiling of 27.75% (or the lenient one without FATAL DISAGREEMENT cases at 59.69%) that the classifiers' performance ought to be compared.

4. A small amount of posts were excluded due to anomalous content and features.

– *Data quality.* This kind of empirical-data crowd sourcing offers practical benefits that cannot be refuted. It can further be argued that, because they reflect real, uncontrolled, and 'untrained' opinions, crowd-sourced sentiment annotations are maximally valid in terms of their naturalness. The downside is naturally that the quality of the resultant annotations may be lower than what can be reached in traditional, more rigorous controlled and vetted annotation campaigns. In particular, the very nature of the Mechanical Turk service is based on the notion of quantity rather than quality as the Turkers expect simple tasks that do not require any comprehensive annotator training as such[5] so that each can be completed in a matter of seconds to reflect the typically paltry per-item pay rates.

Despite these quality concerns, the use of crowd sourcing as a data collection method has proven effective for machine learning in general and sentiment analysis in particular (Hsueh et al. 2009).

Table 1. Distribution of 3-way sentiment judgements from 86 annotators.

POS	NTR	NEG	#	%	Agreement type	#	%
	NTR (3)		46	10.13%	FULL AGREEMENT		
		NEG (3)	45	9.91%	FULL AGREEMENT		
POS (3)			34	7.49%	FULL AGREEMENT		
	NTR (2)		1	0.22%	FULL AGREEMENT		
						126	27.75%
POS (1)	NTR (2)		43	9.47%	NEUTRAL DISAGREEMENT		
	NTR (2)	NEG (1)	40	8.81%	NEUTRAL DISAGREEMENT		
	NTR (1)	NEG (2)	39	8.59%	NEUTRAL DISAGREEMENT		
POS (2)	NTR (1)		23	5.07%	NEUTRAL DISAGREEMENT		
						145	31.94%
POS (1)		NEG (2)	82	18.06%	FATAL DISAGREEMENT		
POS (2)		NEG (1)	56	12.33%	FATAL DISAGREEMENT		
POS (1)	NTR (1)	NEG (1)	44	9.69%	FATAL DISAGREEMENT		
POS (1)		NEG (1)	1	0.22%	FATAL DISAGREEMENT		
						183	40.31%
			454	100.00%			

Hsueh et al. (2009), for example, carried out an analysis of sample-post snippets from the same pool of blog data as ours and confirmed the high quality of ratings from Turkers against those from 'expert' annotators. Turker annotations have also been used to rate Wikipedia articles (Kittur et al. 2008) and news headlines (Snow

5. A typical HIT page contains only simple instructions and/or examples.

et al. 2008): in both studies, the quality[6] of Turker ratings was also comparable to that of expert annotators.

We hence conclude that the seemingly low inter-annotator agreement rates on our data set were not the byproduct of crowd-sourced annotations as such but rather reflect the inherently fuzzy and subjective properties of the underlying sentiment classification task.

3. Related work

Due to the fact that our classification framework involves a complex network of tools, ideas, and phenomena from multiple paradigms, topics, areas, and fields, we cannot provide a full survey here. We hence limit the discussion on the relation of the present study to past work and existing proposals to *political sentiment analysis*.

The recent surge of interest in 'mainstream' (e.g., product/movie review-oriented) Sentiment Analysis and Opinion Mining (Pang and Lee 2008) has touched upon the political domain in the form of a few studies which have followed the standard document topic-classification paradigm and which have simply mapped the default positive-neutral-negative sentiment polarities onto political 'polarities' in bi- or tripartite political systems. These document-level approaches typically use some form of machine learning with no or only shallow linguistic features. Mullen and Malouf (2006) discuss the application of sentiment analysis to informal political discourse to predict political affiliations as RIGHT (Republican, conservative, r-fringe) vs. LEFT (Democrat, liberal, l-fringe) in blog posts using a probabilistic classifier (accuracy ~60.37%). In a similar study, Malouf and Mullen (2006) used web-based Pointwise Mutual Information scoring, supervised machine learning, and citation graph clustering (accuracy 68.48%, ~73%).

Other variants of the same paradigm have focused on classifying public comments on proposed governmental regulations as PRO vs. AGAINST with a combination of sentiment analysis, (sub)topic detection, argument structure analysis, and semantic frame analysis (Kwon et al. 2006). The cultural orientation and ideologies in left- and right-wing political documents were estimated based on co-citation information in Efron (2004) (accuracy ~90%) while congressional floor debates were classified as SUPPORT FOR vs. OPPOSITION TO to (a piece of legislation) using

6. Anecdotal evidence and general opinions amongst the users of crowd-sourced annotations however suggest that their quality depends (sometimes entirely randomly) on the annotation task attempted and the pool of Turkers that participated in it, and that the risk of obtaining junk annotations is always present.

graph-based agreement links between speech segments and speaker identities in Thomas et al. (2006) (accuracy ~70.81%). Others have involved various forms of supervised learning to classify congressional floor debates for general sentiment (Yu et al. 2008) (accuracy ~65.5%); to capture what kinds of subjective perspectives (points of view) are expressed in text pertaining to the ISRAELI vs. PALESTINIAN polar classes (Lin et al. 2006) (accuracy 93.46~99.09% for documents, ~94.93% for sentences); and to classify LEFT-VOICE vs. RIGHT-VOICE blog posts about President Bush's management of the Iraq War (Durant and Smith 2007) (accuracy ~89.77%), amongst others.

Similar approaches can be found in the form of predictive models which include temporal features. An opinion-forecasting approach was described in Lerman et al. (2008) who combine a shallow bag-of-words approach with predefined entities, syntactic parsing, and temporal news coverage models to predict the impact of news on public perception of political candidates. Kim and Hovy (2007) in turn describe a supervised learning system that predicts which party is going to win the election on the basis of opinions posted on an election prediction website (accuracy ~81.68%).

While these studies have reported relatively high 'accuracy' levels, the standard document topic classification paradigm is too coarse to score individual entities. To the best of our knowledge, the role of deeper sentiment analysis in general and fine-grained multi-entity scores in particular has not been investigated fully in the area of political sentiment analysis. The approach closest to ours in this analytical vein, centered around entities, is Van Atteveldt et al. (2008b) who describe a system for automatic analysis of Dutch political newspaper texts that exploits basic semantic sentiment roles such as opinion SOURCES in constructing a basic conceptual semantic network representing a given document. The authors extract semantic relations between political actors derived from syntactic parsing, rule-based mappings between syntactic and semantic roles, political ontologies, a basic anaphora resolution mechanism, and template-based pattern matching to find opinion SOURCE constructions and to determine the general semantic AGENT and PATIENT roles. In the task of determining basic semantic roles at the level of documents, the authors reached 53% (precision): 31% (recall) for opinion SOURCES, 57% (precision): 51% (recall) for AGENTS, and 39% (precision): 36% (recall) for PATIENTS. Instead of classifying entities as such, a related study in Van Atteveldt et al. (2008b) attempted to classify specific positive/neutral/negative relations (e.g., support/criticism, success/failure, evaluation) between political actors and issues in the news coverage of the Dutch 2006 elections. The authors focused on relevant salient words subsumed by subject and object nodes in dependency parse trees as features for machine learning (alongside part-of-speech tag, *n*-gram, and semantic thesaurus class information). An overall F-score of ~63% was reported.

The main difference between these methods and our proposed approach is that the former afford only partial coverage of some pre-specified entities of interest in pre-specified syntactic and semantic positions while a more comprehensive multi-entity scoring framework captures all individual entity mentions and hence offers considerably higher recall.

4. Overview of the classification framework

4.1 Shallow document classification

One of the simplest approaches to text categorisation is arguably the bag-of-words paradigm in which a given text is represented as an unordered collection of independent statistical features pertaining to word or *n*-gram frequencies. Although structural and positional information about words and *n*-grams is discarded altogether, the bag-of-words method is hard to beat in practice. We therefore adopt as a strong baseline classifier a unigram count model (Manning et al. 2008) that operates on stemmed[7] and normalised *n*-grams from the blog posts. Rather than using raw word frequencies directly, we use TF-IDF (Term Frequency-Inverse Document Frequency) vectors to represent the posts themselves. We limited the word vectors to terms that appear in at least 5 different posts, and avoided terms that appeared in over 500 posts. This resulted in 4647 unigram features for classification.

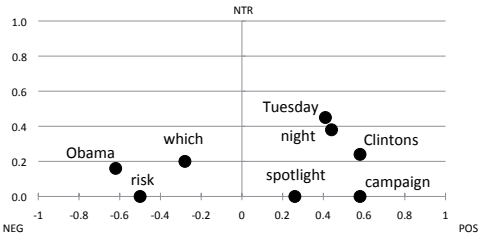
4.2 Deep entity-level sentiment scoring

In order to complement deeper linguistic and sentiment information the holistic evidence offered by the shallow *n*-gram method, we employed a wide-coverage sentiment parser. In particular, we wished to utilise in the analysis information about the overall sentiment expressed in each blog post towards all individual entities mentioned in it. The parser, which is described in greater detail in Moilanen and Pulman (2009), employs compositional sentiment logic, deep grammatical analysis, and large manually compiled sentiment lexica to exhaustively assign sentiment scores to different structural levels across individual words, syntactic phrases, sentences, and documents. In particular, it assigns gradient POS:NTR:NEG sentiment scores for all individual entity mentions (e.g. *"Obama*$^{(+)}$*"*, *"Obama's*$^{(-)}$*"*, *"Barack*$^{(N)}$*"*, *"Chicago*$^{(+)}$*"*, …) and aggregated entity

[7] As output by the standard Porter stemmer (Porter 1980).

topics (e.g. *"Obama"* had 25 mentions 58% of which were positive) in a given blog post. Example (1) shows a sample sentence from our corpus and the gradient sentiment scores that the parser assigned to the [ENTITIES] in it.

(1) *Judging by* [TUESDAY] [NIGHT], *the* [CLINTONS] *would want to share the* [CAMPAIGN] [SPOTLIGHT], [WHICH] *runs the* [RISK] *of making Mr.* [OBAMA] *look weak.*



We transformed the aggregated topical POS:NTR:NEG sentiment percentage counts from all entities mentioned across all 700 posts into unique learning features (e.g., OBAMA_NEG_SCORE, PALIN_NTR_SCORE, IRAQ_NEG_SCORE, …). Example (2) shows the top 25 most frequent topical entities from the posts.

(2) *Obama* (115), *that* (97), *you* (91), *he* (91), *it* (91), *Barack* (88), *I* (84), *they* (81), *we* (81), *who* (79), *what* (70), *McCain* (69), *people* (67), *this* (65), *campaign* (64), *candidate* (59), *there* (56), *president* (53), *John* (52), *time* (51), *election* (51), *all* (49), *one* (48), *comment* (48), *day* (46)

In order to make the features more focused around the key entities and issues in the election, we discarded features from entities that had a frequency of <10 across our corpus. This filtering resulted in 951 entity score features for classification.

4.3 Social network modeling

In addition to serving as an interesting text corpus of political sentiment, our data set contains rich information about the relationships between individual bloggers represented by various hyperlinking structures. As many bloggers link to each other within their posts, such link data can reveal the rich underlying social structures in the political blogosphere. Past research into such linking patterns during the 2004 U.S. presidential election has, for example, shown that bloggers tend to segregate themselves on ideological grounds, with conservative and liberal bloggers separating into tight-knit clusters that have different behavioural characteristics with regards to hyperlinking to other blogs (Adamic and Glance 2005). Such segregation has also been observed in more topically focused blog communities such as war blogs (Tremayne et al. 2006). Analysis of links between ideologically-charged blog clusters similarly shows that links are often used to critique other bloggers (Hargittai et al. 2008).

Since blogs often self-categorise into ideologically-charged clusters, incorporating information about such clusters into our blog-post categorisation model appears intuitively beneficial. We accordingly sought to investigate the possibility of leveraging the above kinds of social linking and sorting phenomena observed in the political blogosphere to facilitate the blog-post classification and labeling task. In this initial study, we take a relatively simple approach to exploring which clusters bloggers find themselves in, with posts then acquiring the features of their parent blogs.

Weakly connected components. While one could look at the individual post-level linking patterns between blogs (e.g., Leskovec et al. 2007), the limited number of post labels to which we currently have access – combined with sparse post-level networks – means that not enough post-level social information may be gleaned from the data. Social networks at the blog level do however provide more information as all hyperlinks observed between blogs in the data set can be aggregated into a directed network. We treated the connections between blogs as unweighted: in particular, if at any point during the year blog A was linked to blog B, then we treated this as a link in the blog graph. The one shortcoming of such an approach is that relationships between bloggers that regularly link to each other are treated the same as one-off links.

Using the *iGraph* package[8] (Czárdi and Nepusz 2006), we determined the location of each blog in the social network, alongside a number of different post-specific properties. In each case, we explored different subgraph types and whether posts were written by a blogger situated in specific subgraphs of the network. For example, a simple subgraph structure within a directed network is a *weakly connected component* which represents subgraphs where all nodes are connected to all other nodes. In our blog network, the largest such weakly-connected component consists of 8,297 blogs, while the second largest has 67. In this case, two variables could be added to the feature vector used to classify a given blog post, namely (1) is the parent blog situated in the largest component, and (2) is the parent blog situated in the second-largest component? From these variables, ten boolean features were generated that indicate which community a given post belongs to, based on which blog was responsible for posting it.

Community detection algorithms. A second plausible approach is afforded by community detection algorithms which aim at finding dense subgraphs within a social network. While a component can be one interpretation of a community within a social network, community-finding algorithms often have more stringent

8. http://igraph.sourceforge.net/, last accessed March 6, 2014.

definitions. The general idea behind many such algorithms is to locate parts of a social network that are more dense than one would expect based on the network as a whole. Such a dense subgraph may imply stronger relationships between members of the subgraph compared to members outside of the subgraph. If we assume that the social network is homophilous by ideology (McPherson et al. 2001) – that is, if blogs tend to link to each other more often when they share similar political views – then we can use their membership within subgraphs as features for classification. Using a *fast greedy community detection* algorithm (Clauset and Newman 2004), a number of communities were detected in the blog network. Fast greedy community detection places nodes into communities in a way that maximises the number of edges within communities, rather than between them. It is an agglomerative approach where each node begins by being within its own community, and communities are then merged to maximise within-community links and minimise between-community links. The ten largest communities found[9] using this approach are outlined in Table 2.

Table 2. The ranks and sizes of the largest communities found in the aggregated blog network.

Rank	1	2	3	4	5	6	7	8	9	10
Size	2077	1659	1131	945	582	484	380	355	98	82

From the above approaches to modeling the social-network structure of our blog sample, 22 cluster features were generated for classification. These come in three major categories:

1. *Two weakly-connected components.* Two of the features track whether the specific blog belongs to one of two of the largest weakly-connected components in the network.
2. *Ten communities* (1). Membership within ten of the largest communities as determined by the fast-greedy community detection algorithm is tracked as the next set of features.
3. *Ten communities* (2). Membership within ten of the largest communities as determined by the leading eigenvector approach to community detection. This is similar to the earlier feature set, but it should be noted that different community detection algorithms (indeed, different random seeds) provide different membership categories.

9. Note that the blog network was symmetrised prior to running the fast-greedy algorithm.

4.4 Overview of algorithms

For all experiments, we used standard Naïve Bayes Multinomial (NBM) and Logistic Regression models available in the WEKA toolkit (Hall et al. 2009), all with their default parameters. By assuming that individual terms appear in a blog post independently of all other terms, the NBM classifier calculates the probability of a given blog post belonging to a positive, neutral, or negative category. This is done by summing up the estimated log-probabilities of individual terms appearing in the categories.

We further made use of 2nd-tier Logistic Regression metaclassifiers which use as their inputs the probabilistic predictions made by three 1st-tier NBM classifiers. As metaclassifiers, we compared two different options, namely (1) a Majority Voting classifier which counts the three class labels predicted by the 1st-tier classifiers and assigns the majority class label to a given blog post, and (2) a Stacking classifier which takes three inputs from the 1st-tier classifiers and treats them as features for the final classification step.

We further experimented with Support Vector Machines (SVMs) and J48 decision tree algorithms. Because they did not perform as well as the above classifiers on our data set, their results are not included in this paper.

5. Experiments

5.1 Experimental conditions

We report the performance of three different feature types (§4) across 5620 features in the following conditions:

1. A Logistic Regression classifier with 22 social network features [SNA]
2. An NBM classifier with 951 sentiment analysis features [SA]
3. An NBM classifier with 4647 unigram bag-of-words features [BOW]
4. An NBM classifier with all 5620 features [ALL]
5. A Stacking classifier with three separate NBM classifiers for (1), (2), and (3) [STACK]
6. A Voting classifier with three separate NBM classifiers for (1), (2), and (3) [VOTE]

Each condition was measured through 10-fold cross-validation which splits the data set into ten different folds (9/10 training vs. 1/10 testing). Each cross-validation run was further seeded with ten different seeds. All reported scores (unless

stated otherwise) represent averages from the 10 × 10-fold cross-validation runs. Three separate baselines are given to reflect a classifier that always outputs a given polarity. The results from a three-way (POS vs. NTR vs. NEG) classification condition are given in Table 3 (below).

Table 3. Average 3-way 10-fold cross-validation results.

	POS	**NTR**	**NEG**
BASELINE ACCURACY (RAW)	27.56	**41.69**	30.75

	SNA (22)	**SA (951)**	**BOW (4647)**	**ALL (5620)**	**STACK (5620)**	**VOTE (5620)**
ACCURACY (RAW)	41.05	46.33	**50.34**	49.70	49.29	49.27
ACCURACY (PAIRWISE)	60.70	64.22	**66.89**	66.47	66.20	66.18
ACCURACY (POS)	68.20	65.88	68.06	69.61	**70.59**	69.20
ACCURACY (NTR)	48.79	55.76	**58.66**	57.08	54.83	56.63
ACCURACY (NEG)	65.10	71.03	**73.96**	72.71	73.17	72.71
PRECISION	39.53	45.85	**50.52**	49.40	49.22	49.43
PRECISION (POS)	37.34	36.32	40.72	**42.87**	42.08	41.92
PRECISION (NTR)	42.69	47.54	**50.32**	48.86	47.31	48.54
PRECISION (NEG)	38.57	53.68	**60.52**	56.48	58.26	57.81
RECALL	37.38	44.35	**48.01**	47.49	45.33	46.47
RECALL (POS)	22.64	31.65	**34.46**	30.58	17.52	30.33
RECALL (NTR)	66.67	58.74	65.41	62.79	**73.50**	67.38
RECALL (NEG)	22.81	42.67	44.15	**49.11**	44.96	41.70
F-SCORE	36.30	44.63	**48.41**	47.72	44.33	46.68
F-SCORE (POS)	28.18	33.82	**37.32**	35.67	24.70	35.18
F-SCORE (NTR)	52.05	52.54	56.88	54.95	**57.56**	56.43
F-SCORE (NEG)	28.66	47.53	51.04	**52.53**	50.73	48.45
SAR	50.43	54.28	**56.87**	56.46	55.81	56.04
KAPPA	8.40	19.62	25.15	**25.52**	22.55	23.38
KRIPPENDORFF	44.56	52.84	54.51	**55.64**	51.38	53.25
PEARSON	12.14	23.35	28.67	**30.81**	28.74	27.80
SPEARMAN	12.12	23.46	28.74	**31.18**	29.37	27.99
FATAL ERRORS	13.15	17.55	16.73	14.69	**10.92**	14.51
GREEDY ERRORS	23.56	32.05	29.03	30.83	**21.78**	26.79
LAZY ERRORS	63.29	**50.39**	54.24	54.48	67.30	58.70

5.2 Evaluation measures

A large number of different evaluation measures can be used to characterise the performance of the classifiers and the features used, each of which highlights a different evaluative aspect.

Accuracy. The first measure set targets the standard notion of 'accuracy' used in traditional factual classification tasks encompassing *Accuracy, Precision, Recall, F-Score,* and *SAR* measures. For these, individual pairwise polarity conditions (POS vs. NOT-POS, NTR vs. NOT-NTR, NEG vs. NOT-NEG) were used. In addition, raw percentage accuracies are reported. Although sentiment interpretation can not be said to be '(in)accurate' in the strictest sense of the term, these measures characterise the overall behaviour of our classifiers in a useful way.

Agreement. The second set of measures focuses on different levels of agreement and correlation between human sentiment judgements and our classifiers by calculating chance-corrected ternary (POS vs. NTR vs. NEG) rates based on the standard Kappa *k*, Pearson's *r* product moment correlation coefficient, Spearman's ρ rank order correlation coefficient, and Krippendorf's α reliability coefficient measures.[10]

Error types. The inter-annotator agreement levels point towards increased ambiguity with NTR polarity due to differing personal degrees of sensitivity towards neutrality/objectivity. Not all classification errors are then equal for classifying a POS case as NTR is more tolerable than classifying it as NEG, for example. We found it useful to characterise three distinct *error classes* or disagreements between human *H* and algorithm *A*. FATAL errors ($H^{(\alpha)}A^{(\neg\alpha)}$ $\alpha \in \{+\ -\}$) are those where the non-neutral polarity is completely wrong: such errors affect the performance of a classifier adversely. GREEDY errors ($H^{(N)}A^{(\alpha)}$ $\alpha \in \{+\ -\}$) are those where the algorithm wrongly made a decision to jump one way or the other, displaying oversensitivity towards non-neutral polarities. LAZY errors ($H^{(\alpha)}A^{(N)}$ $\alpha \in \{+\ -\}$) indicate that the algorithm chose to sit on the fence and displayed oversensitivity towards NTR polarity. We naturally aim at minimising FATAL errors.

10. All accuracy and agreement measures were obtained using R (http://www.r-project.org/) with The built-in correlation functions together with the ROCR (http://cran.r-project.org/web/packages/ROCR/) and IRR (http://cran.r-project.org/web/packages/irr) packages. All pages were last accessed on March 6, 2014.

5.3 Discussion

In absolute terms, the scores are modest. However, when compared to the low human ceiling (27.75~59.69%) (§2.2), they do in fact appear promising. In general, all classifiers easily surpassed the low non-neutral (27.6~30.75%) and neutral[11] (41.69%) baselines. The standalone performance of the social network (SNA) features was not as effective as we expected. In the light of the very small number of features used (only 22), it is in fact surprising that the SNA features worked at all. The average F-score obtained by the SNA features was low (36.3%) mainly due to low recall for non-neutral sentiment. Their pairwise accuracy rates are more favourable as they show that the SNA features are not making random non-neutral predictions. When larger networks are incorporated in the future, social network features can be expected to offer important supporting evidence in the classification task that we are attempting.

Equally promising is the performance of the sentiment analysis (SA) features – especially considering that they only reflect the sentiment scores of a handful of entities and constitute a relatively small feature set (only 951). The underlying compositional sentiment parser, from which the SA features stem, is very sensitive towards non-neutral sentiment which resulted in slightly higher FATAL and GREEDY error rates for the SA features (cf. the lowest LAZY error rate). The SA features are on the whole more balanced than the SNA ones.

Considering the much larger feature set, the performance of the holistic bag-of-words (BOW) features was (perhaps unsurprisingly) very strong. The figures from 5620 features suggest that, as features, unigram evidence still reflects the sentiment properties of a blog post more closely than non-lexical evidence, even when it comes to measuring sentiment towards or about a single entity. The BOW features were behaviourally closer to the SA features than to the SNA ones which may be due to the fact that the SA entity features latently represent salient unigrams (e.g., *"Obama"*).

Our hypothesis concerning the leveraging power of the SNA and SA features was confirmed partially as small gains over the BOW model were obtained in some conditions by using the entire feature set (ALL) for a single classifier. This can in particular be seen in the higher agreement/correlation rates and lower amounts of FATAL errors. Stacking and voting amongst the three individual NBM classifiers provided further boosts in some conditions.

11. With the exception of the SNA features.

Interestingly, all classifiers displayed a general tendency towards faring worse with positive polarity than with negative polarity, especially in precision and recall. All classifiers reached their highest recall and F-scores with neutral polarity. Moreover, the amount of FATAL errors stayed below 17% throughout which suggests that all feature types are generally pointing at the right direction. This correlates with the fact that a full 31.94% of the annotations involved conflicting annotations around neutral polarity (NEUTRAL DISAGREEMENT). The classifiers can accordingly be expected cope better with non-neutral sentiment. When all neutral cases were excluded from the evaluation, a small subset of 90…145 non-neutral test cases was examined in order to verify this. Although the subset is too small to draw any definite conclusions, it can nevertheless shed light on how the classifiers did actually do with the core non-neutral cases which are arguably more important for a sentiment classifier than neutral, objective cases.

Table 4 confirms the expected complications caused by neutral polarity in that non-neutral precision scores go as high as 75.68%. Interestingly, the boost over the BOW features given by the non-lexical SNA and SA features is much clearer in the 2-way condition as the combined ALL features and the STACK and VOTE classifiers all outperformed the BOW classifier. These 2-way scores seem to confirm that the SNA and SA features can indeed be used to leverage shallow unigram features in non-neutral cases.

Lastly, we looked at the most informative features based on the Chi-squared and information gain measures across all 5,620 features. Of the top 50 features ranked by the two measures, 29 were entity scores from the SA feature set, with the negative entity score for *"Obama"*, *"Chicago"*, and the positive score for *"rhetoric"* topping the ranks alongside high-ranking unigrams such as *"Wright"*, *"pastor"*, and *"Rezko"*, amongst others. This further confirms the utility of the entity-level SA features for our classification task. The SNA features did not rank high as features, however.

Table 4. Average 2-way 10-fold cross-validation results.

	SNA (22)	**SA (951)**	**BOW (4647)**	**ALL (5620)**	**STACK (5620)**	**VOTE (5620)**
ACCURACY (PAIRWISE)	63.16	69.84	73.55	76.14	**77.13**	74.23
PRECISION	63.13	69.14	73.06	75.52	**75.68**	73.51
RECALL	63.10	69.73	73.50	**74.66**	72.51	73.78
F-SCORE	63.07	69.21	73.10	**74.95**	73.42	73.57
SAR	55.20	61.58	65.24	**67.32**	67.29	65.77
KAPPA	26.19	38.62	46.35	**49.99**	47.25	47.19
PEARSON	26.23	38.87	46.55	**50.17**	48.08	47.28

5.4 Significance of results

What is particularly interesting about the results above is the implied relationship between political ideologies and the social structure of the political blogs themselves. Not only do these results imply that one can improve political text mining through structural features of the system being analysed, but these results also imply a relationship between social structure and political discourse within the social system itself. Indeed, the results imply that simply knowing the social structure (i.e. link relationships between bloggers) is enough to surmise the general political stance of many bloggers.

This is particularly important for research into political discourse, as the 2004 and 2008 elections experienced a large amount of political blogging, and social media as a whole played a key role in U.S. political discourse during this time. Since 2008 (when the data was collected), new forms of social media (e.g., Twitter, Pinterest, etc.) have begun to pervade the political social media landscape. As these social technologies become engrained in our society – and as such, into popular political discourse – it is important to understand the interplay between the social structure of those participating, and the results of political discourse taking place within these communities. If the social structure alone can effectively predict the political views of those participating in the discourse, we must strive to understand the social and political reasons behind these results.

We encourage researchers – computational linguists, machine learning experts, and political scientists – to take these results and explore whether similar relationships between structure and discourse appear in other types of communities and social systems (be they online or not). While the future work we propose in Section 5.5 focuses tactically on extending the classification accuracy of our models, the findings also suggest broader applications to the fields of political science and opinion research.

5.5 Future work

The proposed classification framework and the results obtained in this initial study open up many avenues for future work.

Sentiment analysis. Although their utility is intuitively appealing, it is unclear how far the capabilities of current deep-sentiment analysis tools could bring the analysis considering the subjective nature of the classification task. The most obvious avenue for improvement is to incorporate semantic sentiment-role information (cf. Van Atteveldt et al. 2008a) alongside information about basic Named Entities

in order to have access to features expressing people, places, and organisations but also to richer political domain knowledge in the form of ontologies and other evidence. Regarding the deep general-purpose sentiment parser that we employed, further improvements can arguably be made by tuning its underlying lexicon towards the political domain. The entity-level sentiment scores that were used in this study can further be boosted by resolving pronoun mentions to their antecedents (cf. Van Atteveldt et al. 2008a) which, judged by the number of pronouns amongst the top-scoring topical entities in Example (2), can be expected to have a tangible impact on the classifiers.

Another useful research question is whether longer bi- and tri-grams could improve accuracy beyond what we would expect from further blog corpora. Although longer *n*-grams can rudimentarily model some further linguistic features of political discourse, past research in document-level sentiment analysis suggests that simple binary unigram presence features suffice.

Clustering by topic. Another approach to analysing political blogs comes in the form of blending social-network analysis with entity extraction. For every post in our corpus, we have a list of entities that are mentioned in the post, together with their sentiment labels. It is then possible to extract all individual entities mentioned in each post and build a bipartite network representing the data set. In this case, we have a matrix with columns representing entities and rows representing the individual posts (e.g., a value of 1 at the i^{th} row and j^{th} column of the matrix representation of the network denotes that the i^{th} post mentioned the j^{th} entity). That representation would allow us to observe topic overlaps between different posts which would be analysed by constructing a social network between posts, where an edge between posts exists if the posts have some amount (e.g., 10) of entities in common with each other. Such a inter-document network can help elucidate clusters and communities based on topics – not wholly unlike how we use community detection algorithms as part of the social network analysis to locate groups of bloggers that tend to communicate with each other. In this case, however, sharing large numbers of topics or entities in discussions may mean that specific topics – or at the very least, having topics in common – lead to similar sentiment scores. Unfortunately, carrying out such an analysis at this stage yields very poor results for topic clusters: it appears that most blog posts have a large number of entities in common which causes standard community finding algorithms to group the entire set of posts together as one large community. We conjecture that this is an artifact of the selection method used for blog posts discussing Barack Obama. At this stage, the topical clustering method still merits further research.

Advanced social network analysis. At this stage, only a few algorithms have been investigated to combine the various feature sets. While we could explore richer features based on entity extraction, *n*-grams, and deeper social-network analysis discussed above, we would still be developing feature vectors for each case and build classifiers that operate on them. Since the algorithms in Weka are not optimised as such for dealing with social network analysis, one extension to our research is to build algorithms that directly integrate predictions into a social network, rather than using features that reflect the current 'membership-within-community' approach. One can, in particular, use the outputs of the entity-level sentiment scores or unigram probability estimates and apply them as labels in the social network. The social network itself could then be used to reinforce the labels and see which ones appear realistic based on previous observations.

Dataset. A major challenge with the current initial study was data sparsity. Only 700 posts out of 2.8 million were labelled, making it very difficult to extract useful information from the social network features in the data set. We hypothesise that, as more posts are labelled and one gets a finer-grained picture of how bloggers self-organise themselves into topical and political communities, social network features should become more relevant and important. Another issue that ought to be investigated is the concept of data-set shift. The fact that many terms and phrases in the political domain develop and change their (non-)affective connotations over time is a key challenge for any text-based political blog and sentiment analysis framework (especially for shallow classification methods). For example, a term such as *"Alaska"* may have had different affective connotations before and after Sarah Palin was announced as the Republican vice presidential candidate.

Opinions and ideology. The experiments in this paper focus on extracting sentiment toward a specific politician or entity. A more difficult challenge would be extracting the ideology of a post or sentence, as political ideologies tend to be more fluid, rhetorical, and generally difficult to classify even by humans. An extension of the work above would be to train models for political ideology (e.g., liberal versus conservative), rather than positive or negative sentiment. Such work has already been attempted in analyzing political blogs (Durant and Smith 2006), but merits further rigorous research. An approach similar to the one taken in this paper would be ideal.

6. Conclusion

This chapter targets an entity-centric document-level sentiment classification of political blogs. We presented the results of an initial study that sought to investigate the feasibility of combining linguistic indicators of political sentiment with non-linguistic information obtained from social network analysis. Using crowd-sourced sentiment annotations centered around Barack Obama during the 2008 U.S. presidential election sampled from a large corpus of 2.8 million blog posts, a hybrid machine-learning and logic-based framework was employed which relies on standard shallow document classification, deep linguistic multi-entity sentiment analysis and scoring, and social-network modeling. The initial results demonstrate the complexity of the task, and point towards the positive effects of learning features that exploit entity-level sentiment scores and social network structure.

Acknowledgements

We would like to thank the *Predictive Modeling Group* at IBM for providing us with the blog data, and Nigel Crook at Oxford University Computing Laboratory for supporting code.

References

Adamic, L. and N. Glance. 2005. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery (LinkKDD 2005,* New York, NY), pp. 36–43.

Clauset, A., M. E. J. Newman and C. Moore. 2004. Finding community structure in very large networks. *Physical Review E*, 70(6), 06611, 6 pages.

Csa´rdi, G. and T. Nepusz. 2006. The igraph software package for complex network research. *International Journal Complex Systems*, 1695.

Durant, K. T. and M. D. Smith. 2006. Mining sentiment classification from political web logs. In *Pro-ceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (WebKDD-2006)*.

Durant, K. T. and M. D. Smith. 2007. Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. In Advances in Web Mining and Web Usage Analysis: *Proceedings of the 8th International Workshop on Knowledge Discovery on the Web (WEBKDD 2006*, Philadelphia, PA), pp. 187–206.

Efron, M. 2004. Cultural orientation: Classifying subjective documents by co-citation analysis. In *Style and Meaning in Language, Art, Music, and Design: Papers from the 2004 AAAI Fall Symposium*, Arlington, VA. Technical Report FS-04-07, pp. 41–48.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten. 2009. The weka data mining software: An update. *ACM SIGKDD Explorations Newsletter* 11(1), pp. 10–18. DOI: 10.1145/1656274.1656278

Hargittai, E., J. Gallo and M. Kane. 2008. Cross-ideological discussions among conservative and liberal bloggers. *Public Choice* 134(1), pp. 67–86.

Hsueh, P. Y., P. Melville and V. Sindhwani. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pp. 27–35. DOI: 10.3115/1564131.1564137

Kim, S.-M. and E. Hovy. 2007. Crystal: Analyzing predictive opinions on the web. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL*, Prague), pp. 1056–1064.

Kittur, A., E. H. Chi and B. Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the 26th annual SIGCHI conference on Human factors in computing systems*, Florence, pp. 453–456.

Kwon, N., S. W. Shulman and E. Hovy. 2006. Multidimensional text analysis for erule-making. In *Proceedings of the 7th Annual International Conference on Digital Government Research (DG.O 2006*, San Diego, CA), pp. 157–166.

Lerman, K., A. Gilder, M. Dredze and F. Pereira. 2008. Reading the markets: Forecasting public opinion of political candidates by news analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008*, Manchester), pp. 473–480.

Leskovec, J., M. McGlohon, C. Faloutsos, N. Glance and M. Hurst. 2007. Cascading behavior in large blog graphs: Patterns and a model. *Society of Industrial and Applied Mathematics: Data Mining (SDM07*, Minneanapolis, MN). Tech report (12 pp.): CMU-ML-06-113.

Lin, W. H., T. Wilson, J. Wiebe and A. Hauptmann. 2006. Which side are you on? identifying perspectives at the document and sentence levels. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X* New York, NY), pp. 109–116.

Malouf, R. and T. Mullen. 2008. Taking sides: User classification for informal online political discourse. *Internet Research* 18(2), pp. 177–190. DOI: 10.1108/10662240810862239

Manning, C. D., P. Raghavan and H. Schütze. 2008. *An Introduction to Information Retrieval*. Cambridge: CUP. DOI: 10.1017/CBO9780511809071

McPherson, M., L. Smith-Lovin and J. M. Cook. 2001. Birds of a feather: Homophily in social net- works. *Annual Review of Sociology*, 27(1), pp. 415–444. DOI: 10.1146/annurev.soc.27.1.415

Moilanen, K. and S. Pulman. Multi-entity sentiment scoring. 2009 In *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2009*, Borovets, Bulgaria), pp. 258–263.

Mullen, T. and R. Malouf. 2006. Preliminary investigation into sentiment analysis of informal political discourse. In *Computational Approaches to Analyzing Weblogs: Papers from 2006 AAAI Spring Symposium* (Stanford, CA), pp. 159–162.

Pang, B. and L. Lee. 2008. *Opinion Mining and Sentiment Analysis, volume 2 of Foundations and Trends in Information Retrieval*. Now Publishers.

Porter, M. 1980 An algorithm for suffix stripping. *Program* 14(1), p 3.

Snow, R., B. O'Connor, D. Jurafsky and A. Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008*, Honolulu, HI), pp. 254–263.

Thomas, M., B. Pang and L. Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2006*, Sydney, Australia), pp. 327–335.

Tremayne, M., N. Zheng, J. K. Lee and J. Jeong. 2006. Issue publics on the web: Applying network theory to the war blogosphere. *Journal of Computer-Mediated Communication* 12(1), pp. 290– 310. DOI: 10.1111/j.1083-6101.2006.00326.x

Van Atteveldt, W., J. Kleinnijenhuis and N. Ruigrok. 2008a. Parsing, semantic networks, and political authority using syntactic analysis to extract semantic relations from Dutch newspaper articles. *Political Science* 16(4), pp. 428–446.

Van Atteveldt, W., J. Kleinnijenhuis, N. Ruigrok and S. Schlobach. 2008b. Good news or bad news? Conducting sentiment analysis on Dutch text to distinguish between positive and negative relations. *Journal of Information Technology & Politics* 5(1), pp. 73–94. DOI: 10.1080/19331680802154145

Yu, B., S. Kaufmann and D. Diermeier. 2008. Exploring the characteristics of opinion expressions for political opinion classification. In *Proceedings of the 9th Annual International Conference on Digital Government Research, Partnerships for Public Innovation, Vol. 289 (DG.O 2008*, Montreal), pp. 82–89.

CHAPTER 4

Issue framing and language use in the Swedish blogosphere

Changing notions of the outsider concept

Stefan Dahlberg and Magnus Sahlgren
Department of Political Science, University of Gothenburg
and Gavagai, Stockholm

Issue framing has become one of the most important means of elite influence on public opinion. In this paper, we introduce a method for investigating issue framing based on statistic analysis of large samples of language use. Our method uses a technique called Random Indexing (RI), which enables us to extract semantic and associative relations to any target concept of interest, based on co-occurrence statistics collected from large samples of relevant language use. As a first test and evaluation of our proposed method, we apply RI to a large collection of Swedish blog data and extract semantic relations relating to our target concept "outsiders". This concept is widely used in the public debate both in relation to labour market issues and socially related issues.

Introduction[1]

Issue framing has become one of the most important means of elite influence on public opinion. According to prior work, we understand issue framing as a process where a communicator "defines and constructs a political issue or public controversy by emphasizing a subset of potentially relevant considerations and *[thereby pointing to the*] essence of the issue" (Slothuus 2008; Slothuus and de Vreese 2008). However, we still have a limited understanding of how arguments and rhetoric from political parties actually influence the formation of political opinion. We know that elites attempt to influence opinions by framing issues, i.e., by presenting alternative descriptions and interpretations of different issues. A great body

1. The authors are in great debt to Rebecka Åsbrink and Henrik Lindholm at the Department of Political Science, University of Gothenburg, for their assistance with collecting and analyzing the party related documents that form the basis for the qualitative part of this study.

of literature has also demonstrated the impact of issue frames on public opinion (Chong and Druckman 2007; Kinder 2003).

We believe that an important key towards a deeper understanding of framing theory is to view framing as a dynamic process in which issues are continuously refocused and redefined by *language use*. Parties and commentators are actively framing issues by the very words they choose to speak about it; speaking about "X" in terms of "Y" can send a completely different message than speaking about it in terms of "Z". It is well known that language use may influence the outcome of policy debates; the chances of obtaining one's goals can be dramatically improved by getting everyone to debate an issue "in your terms" (Naurin et al. 2009). The goal of any skilled advocate is to get her idea to catch on, to reach the tipping point where her way of thinking is not just one way of thinking, but *the* way of thinking.

In this chapter, we introduce a method for investigating issue framing based on statistical analysis of large samples of language use. Our method uses a technique called Random Indexing (RI), which enables us to extract semantic and associative relations to any target concept of interest, based on co-occurrence statistics collected from very large samples of relevant language use. If "X" and "Y" both occur together with "Z", our method will relate "X" and "Y". We suggest that such semantic relations are indicative – if not constitutive – of framing, and that this type of quantitative analysis therefore is an attractive method for investigating issue framing.

As a first test of our proposed method, we have applied RI to a large collection of Swedish blog data from the period 2008 to 2010, and extracted semantic relations to a target concept referring to the notion 'outsiders' (*utanförskap* in Swedish), which is a concept that has been widely used in the public debate by parties, commentators, etc., especially since the national election of 2006. In the public debate, the concept is often used in relation to labour-market issues but also to socially related issues. This paper exemplifies and discusses the various types of relations, and their implications within issue-framing theory. The analysis starts with a qualitative/quantitative approach where we have been tracking the language use for the concept of 'outsiders' among the two dominant and opposing parties in Swedish politics, the Social Democratic Party and the Conservative Moderate Party. The results from this part of the study are then used as a bench-mark for a deepening understanding of the framing of the issue in the blogosphere.

The case of Sweden: Issue framing and the 'outsider' concept

For several decades the Social Democratic Party has been dominant in Swedish politics. With only a few interruptions, in 1976 and in 1991, the Social Democratic Party has been governing with support from the former Communist Left Party and the Green Environmentalist Party. However, during the election of 2006 the chain was broken when the Conservative Moderate Party together with the Center Party, the Liberal Party and the Christian Democratic Party formed an alliance of bourgeoisie parties. With a common platform, the bourgeoisie parties succeeded in breaking the Social Democratic dominance. For the first time in Swedish modern history, they also managed to be reelected for government in 2010. One, amongst many, explanations for the success of the bourgeoisie parties is that they managed to take over the ownership of the employment issue, which is an issue that traditionally has been intimately connected with the Social Democratic Party (Oscarsson and Holmberg 2008).

The literature offers few theoretical explanations for how such a sudden shift in issue ownership could occur (see Walgrave and De Swert 2007 for a review of the subject). In this respect, we believe that policies and issues should not be thought of as being fixed but rather, they are dynamic in the sense that an issue is defined by how parties and commentators speak about it. The language use in the public debate might thus affect voters' perceptions of parties' issue ownership; e.g., which party is perceived as being most competent to handle the issues at stake.

In order to evaluate the applicability and functionality of the RI-model for the purpose of text analysis in general and framing theory in particular, we have tried to identify policy issues and domains where we can expect different language use or frames to be present. We will focus on a much debated issue in Swedish politics during recent years known as 'outsiders'.

The term 'outsiders' has been present in Swedish politics for some ten years, in which period it was used in relation to different forms of outsiderness/alienation.[2] However, during the Swedish national election in 2006, the term was reintroduced as a targeted concept by the alliance of bourgeoisie parties, who turned it into a prioritized goal to decrease the number of 'outsiders' in Swedish society. Their

2. The term 'outsiders' may thus, in general terms, refer to situations where individuals or groups of individuals are or are experiencing that they are excluded from a group or community that is perceived as being desirable to be a part of. The concept may thus pertain to groups of varying sizes that are not accepted in a society, such as different forms of minority groups that are not reckoned or accepted as belonging to the 'normal' society, which in turn will undermine their opportunities to participate to the same extent as the majority population.

definition of the concept was now referring to 'broad unemployment'.[3] This link between 'outsiders' and 'broad unemployment' has in turn been heavily criticized, especially among opposition parties.[4] The critique has mainly been about matters such as which segments of the population should be classified into the category of broad unemployment and whom should not.

In the academic literature, the 'outsider' concept often refers to unemployment and pertains to groups of citizens outside the labor market (see fc. Rueda 2005; Rueda 2006) and this is how the bourgeoisie parties have been referring to the concept. Nevertheless, the term 'outsider' remains a nuanced and imprecise term and the labor market is only one arena for 'outsiderness'. One can in this respect also speak about social, cultural or political outsiders, as examples of 'outsiderness'. This has also been the case in Swedish politics, where it traditionally also has been used in relation to issues of segregation and integration.

The fact that the concept has been used in various contexts at the same time, and the fact that the bourgeoisie parties never gave any clear definition when they reintroduced the term 'outsider', makes the notion imprecise and multiplex. This makes the term, or rather the usage of the term 'outsider' in the Swedish public debate, a fruitful notion for analyzing framing effects. The hypothesis is that the 'outsider' concept is used and spoken about in different terms by different actors and in different contexts. How exactly the concept is used and understood in the public debate is, of course, an empirical question.

Methodological considerations

Our analysis of how the 'outsider' concept is framed consists of both a qualitative/quantitative and a purely quantitative part. The purpose of the qualitative/quantitative analysis is to create a background and a benchmark for how the two main political alternatives in Swedish politics frame 'outsiders'. For this part of the analysis we gathered official documents, articles, news-articles and speeches from the parties' official web-pages. For the Social Democratic Party, a total of 195 documents in which the term 'outsiders' is explicitly mentioned was found during

3. Broad unemployment is in turn defined as the part of the population between 16–64 years of age minus the amount of gainfully employed and minus the amount of students that were not seeking employment, divided by the entire population within 16–64 years of age (source: http://www.slideshare.net/guest3fc6c8/utanforskap-2009-moderaternas-egen-definition).

4. See fc. http://www.svd.se/opinion/brannpunkt/thomas-ostros-utanforskapet-vaxer_4459805.svd

the time-span from 2008 and onwards. For the Conservative Moderate Party, the corresponding number was 148 documents from 2001 to present day.

The qualitative analysis was done in two steps. First, we aimed for an unbiased skim through the documents in order to obtain knowledge about the contexts in which the word 'outsider' appears. From the public debate we knew that we could expect the 'outsider' concept to be used in relation to the labor market; however, in common lexical definitions it can also be related to culture, language and geographical segregation. We have tried to be as explicit as possible by including snippets directly from the texts in order to illustrate our findings and by listing the original Swedish snippets in the Appendix. The second step was to quantify the revealed relationships in order to get a more systematic understanding of the contexts in which the 'outsider' concept appears most frequently.

The second, quantitative, part of the study applies RI to a large data set of Swedish blogs in order to extract semantic relations to the 'outsider' concept. These semantic relations provide information about the words that are most frequently related with the 'outsider' concept in the blogosphere – i.e., how the issue has been framed in this particular data set. This gives us an indication of how the concept is framed in the current debate, and how the frames are shifting in the blogosphere. However, such analysis will not by itself be able to automatically tell or determine to what extent these frames are salient in the blogosphere nor whether these frames are directly related to the frames used by any of the political parties. In order to establish this, we need to manually compare the RI similarities to the benchmark analysis.

Random Indexing

Random Indexing (RI; Kanerva et al. 2000; Sahlgren 2005) is a statistical text-analysis technique that can be applied to massive amounts of text data in order to extract semantic and associative relations between words. The RI technique is a specific implementation of a *word-space model* (Schütze 1993), which is a breed of computational semantic models that use co-occurrence statistics to compute similarity between words. These models represent words by high-dimensional *context vectors*, such that each dimension represents a particular context, and each element indicates the (normalized) frequency of occurrence of the word in that particular context. This means that words that have occurred in similar contexts get similar context vectors, and that the context vectors therefore can be used to compute similarity between words using standard vector similarity metrics. These similarities are interpreted as indicating semantic or associative relations, depending on what type of context is used in collecting the occurrence

information; if word types are used as context, the model will extract semantic similarities, while if documents are used as context, it will extract associative similarities (Sahlgren 2006).

As an example of how word spaces can be used to investigate language use, imagine we are interested in how the word 'labour' has been used in a large collection of texts on political issues. Applying a word-space model on this data can tell us both which words are used in similar ways (e.g. 'employment' and 'job') and which words have an associative relation with 'labour' in this particular data (e.g. 'welfare' and 'security'). Together these analyses give us a good understanding of the usage of words, and it allows us to find text-specific similarities and associations that will not be present in standard lexical or conceptual resources. This is particularly useful when dealing with very productive and dynamic text styles like those typically encountered in social media and the blogosphere.

Word-space modeling has become a standard method in natural language processing for capturing word usage, and models have proven their mettle in an impressive range of large-scale linguistic learning tasks and text analysis applications, including automatic thesaurus construction, terminology mining, word categorization, word sense disambiguation, document clustering, knowledge assessment, text categorization, information retrieval, and modeling of various behavioral effects in psycholinguistic and cognitive experiments (see, e.g., Turney and Pantel [2010] for an overview of NLP methods and applications). However, word-space models tend to suffer from scalability and efficiency issues due to the heavy algebraic machinery involved (e.g., issues regarding very high dimensionality, or the use of matrix decomposition techniques). RI, on the other hand, was developed specifically to overcome problems with scalability and efficiency when dealing with high-dimensional data. Instead of representing each context as a separate dimension, which inevitably leads to very high-dimensional models that are susceptible to both scalability and efficiency issues, RI uses fixed-dimensional vectors in which each context is represented by a small number of randomly chosen dimensions. Every time a word occurs in a context, *all* the elements representing that context is incremented in the word's fixed-dimensional context vector. This use of distributed representations obliterates the need for dimension reduction, and ensures that the dimensionality of the context vectors never increases, even if the data continuously does. We refer to Sahlgren (2005) for a more thorough introduction to RI.

As mentioned above, our evaluation of the applicability of RI for investigating framing effects is operationalized in several steps. The first step consists of a qualitative study of how the term 'outsiders' is used and spoken about on the official party websites of the two largest and opposing parties: the Social Democrats and the Conservative Moderate Party. The second step applies RI to a 1.5 billion-word database of Swedish blog texts, collected between November 2008 and September

2010, and provided by the Swedish blog search engine Twingly (www.twingly.com). The purpose of this step is to investigate which words are mostly associated with 'outsider' in the public blogosphere. We also try to identify which of the parties' language use is most dominant in the blogosphere.

The fact that we only use data from blogs and from the parties' web-pages implies that the results should not be considered to be valid for the general public debate as such. The main purpose with this specific study is not to maximize the external validity, but rather to evaluate the prospects for using RI for investigating framing effects empirically on large amounts of data. By focusing on the blogosphere we have the opportunity to do this with data that are easily accessible.

Language use by the Social Democratic and the Conservative Moderate Party in relation to 'outsiders'[5]

The Conservative Moderate Party

The Conservative Moderate Party is the largest of the bourgeoisie parties in Swedish politics and since the election of 2006 they have formed the government in an alliance with the three remaining bourgeoisie parties (the Center Party, the Liberal Party and the Christian Democratic Party).

The results from our qualitative analysis indicate that for the Conservative Moderate Party, the word 'outsiders' seems to be synonymous with unemployment. In documents, newsletters and speeches published on their webpages, the word 'outsiders' generated a total of 148 hits during a time span between fall of 2008 to September 2010. The word appears together with 'unemployment' (Examples 1 and 2), 'early-retirement pension' (Example 3), 'the opposition', 'health insurance' and 'disability pension':

(1) The goal with the change in the system is to help people to get back into employment and decrease the gap between outsiders and employment.[6]

(2) We choose to carry out understandable policies for full employment and fewer outsiders.[7]

---

5. The quotes in this section are translated into English to facilitate readability. The original Swedish text snippets are listed in the Appendix.

6. Article published 2010-03-01. http://www.moderat.se/web/Kortare_vagar_tillbaka.aspx

7. Published 2009-09-15. http://www.moderat.se/web/Fokus_pa_jobb_och_valfard_i_regerings forklaringen_1.aspx

(3) During the Social Democrats' time in power there was a dramatic increase in sick leave. To improve the stats they transferred 140 persons each day into early retirement pension and thereby labeled them as permanent outsiders.[8]

Of the publications that mention 'outsiders' a vast majority focuses on a group that is the opposite of employed, while a few publications aim to fuse it with the opposition. A distinction in this respect is that the news and newsletters lean more towards the unemployment aspect while the speeches are more focused on the opposition.

Only a hand full of the documents deviate from the use of the term 'outsiders' in relation to unemployment. In some examples the term 'outsiders' is spoken about in national contexts, as opposed to an EU membership context, 'outsider' as the opposite of membership of the EMU. In a few instances where the term 'outsider' is used in relation to other issues, such as schools, fighting crime, senior citizen retired out of old age and issues concerning segregation and integration, it is still used as an indicator of those who are not employed. On the topic of criminality among under aged, Example (4) is taken from an article by Jan R Andersson and Krister Hammarbergh, both members of the Conservative Moderate Party and of the Swedish Parliament, published in *Västerviks-Tidningen, Vimmeby Tidning, Oskarhamns-Tidningen* and *Kinda Posten*.

(4) It is necessary to see what causes criminality: often it is unemployment, lack of social fellowship and lack of security. This is why we need to work even harder to decrease the number of outsiders: higher employment, new jobs, better school and to integrate the immigrants and refugees. Through these actions, we can provide for more children to have a secure childhood – the most efficient preventive action against criminality. By working for a society with fewer outsiders we also work against crime. This is a connection so strong, that it sometimes is forgotten.[9]

Another Example (5) of how 'outsiders' are used in relation to crime comes from the questions and answers section of the webpage.

(5) Crime is fought by a strong justice system and policies for fewer outsiders and less unemployment.[10]

Cristina Husmark Pehrsson, Minister of Welfare and the minister responsible for Nordic cooperation, uses the term 'outsiders' when discussing the economic terms for senior citizens (Example 6).

---

8. Article published 2009-10-26. http://www.moderat.se/web/Fran_sjukskriven_till_arbete_1.aspx

9. http://www.moderat.se/web/Nyheter_1565.aspx

10. http://www.moderat.se/web/2a8a2a45-c5ba-466e-b42d-0d0fcac07267_2_1_1.aspx

(6) A high number of outsiders threatens the welfare since fewer contribute to create the limited resources that have to be shared by more people. The government's efforts for high employment are the only long term sustainable way to protect the welfare. (---) The measures that have been taken to lower the number of outsiders also contribute to secure the pensions.[11]

As these quotes suggest, even though being used in various contexts, the term 'outsiders' mainly refers to people who for various reasons are outside the labor market. In several cases this connection is implicit; get 'outsiders' into employment and they will cease being 'outsiders'. Concluding, the Conservative Moderate Party in general is fairly unambiguous in their use of the term 'outsiders'. Mainly it is referred to in terms of employment.

The Social Democratic Party

During the time-span between 2008 until today, the Social Democratic Party is mainly using the word 'outsiders' in a similar manner as the Conservative Moderate Party does. However, an interesting finding, when stretching focus some more years back in time, is that the party earlier tended to use the concept in a row of shifting contexts. For example, before the parliamentary election 2002 the Social Democrats mentioned the concept of 'outsiders' in their election pamphlet about integration and diversity.[12] Mainly, the party considered the word 'outsiders' connected to immigration, segregation and discrimination during this period.

Four years later, the party still talked about 'outsiders' in connection to immigration, segregation and discrimination.[13] But in 2006, the Social Democrats also talked about 'outsiders' in connection with crime and safety.[14]

In 2007 the party starts to turn the phrase. Like the Conservative Moderate Party, the Social Democrats have begun to mention 'outsiders' in connection with employment and joblessness. But, only when referring to the definition of 'outsiders' as a creation of the Conservatives and with a view to criticize the right-wing government. In her opening speech introducing the party leader debate of the

---

11. http://www.moderat.se/web/Fler_i_arbete_raddar_pensioner.aspx

12. http://www.socialdemokraterna.se/upload/val/val_02/integration.pdf

13. http://www.socialdemokraterna.se/Vart-parti/Socialdemokratiska-riksdagsgruppen/Luciano-Astudillo/nyheter/Arkiv/Artiklar/Artiklar-arkiv/Artiklar-riksdagsaret-200607/Integrationspolitik-utan-tanke-/

14. http://www.socialdemokraterna.se/upload/val/val_06/valblad06/valblad_060828/valblad%205_trygghet%20mot%20brott_060821.pdf

13th of June 2007, Mona Sahlin, party leader of the Social Democrats criticized the government for cutting the unemployment benefit fund.[15]

In the summer of 2008, the concept of 'outsiders' became more and more associated with jobs, welfare, and safety. Here is a speech by Mona Sahlin in Vitabergsparken, Stockholm:

(7) What is the answer from the government Reinfeldt today? In June the outsiders increased for the third month (---) It's been half a term of office. It's time for the evaluation. It is time for proofs. Have they done what they intended to? What about the jobs? What about the outsiders? What about the welfare and safety?[16]

One year later, during the summer of 2009, the adoption of the definition is completed. The Social Democrats now speak of 'outsiders' and employment or unemployment. Often without referring to the definition as a creation of the conservatives (Example 8, Mona Sahlin during the week of politics in Almedalen, Sweden):[17]

(8) The employment rate must increase – among the youth, among immigrated Swedes, among part-time working women, and, never the least, amongst the elderly. To reach that goal the absence of freedom called joblessness and outsiderness must be defeated – and for that matter, the involuntary part-time unemployment. If we want to get there, the concept of the "working line" must apply to everybody.

On the 29th of October 2009, Mona Sahlin wrote in an article on the official homepage (Example 9).[18]

(9) During the last year, the numbers of unemployed have increased with 100 000. Since the election in 2006 the outsiders increased with 70 000 (---) For the election of 2010, the joblessness will be the most important issue.

---

15. http://www.socialdemokraterna.se/upload/MonaSahlin/Dokument/Mona%20Sahlins%20anf%c3%b6rande%20i%20riksdagens%20partiledardebatt.pdf

16. http://www.socialdemokraterna.se/Mona-Sahlin/Tal/2008/Mona-Sahlins-sommartal-i-Vitabergsparken-2008/

17. http://www.socialdemokraterna.se/Mona-Sahlin/Tal/2009/Mona-Sahlin-anforande-Almedalen-2009/

18. http://www.socialdemokraterna.se/media/pressarkivet/nyhetsarkivet-2001--/mona-sahlin-jobben-ar-viktigast/

However, when the Social Democrats want to talk about perceived consequences of the politics formed by the right-wing alliance, they focussed on the social angle of 'outsiderness'. Example (10) is a cut from an article by eight politicians, six from the Social Democratic Party, one from the Greens and one from the Left Party. The article was published in the Swedish newspaper *Corren* on the 26th of February 2010.[19]

(10) Four years ago, the right-wing parties won the election on their promise of more employment and decreasing outsiderness. Unfortunately, when we now evaluate the term of office, the result is depressing. The government throw out tens of thousands Swedish citizens from health insurance to a reality with no insurance at all. In just six months the numbers of unemployed have increased with more than 100 000 persons, and numbers of outsiders with 70 000. Receivers of public assistance are calculated to increase with 50 percent between 2006 and 2011.

And Example (11) is another snippet of news from the Social Democratic homepage, published on the 4th of March 2010.[20]

(11) Today, statistics from Statistics Sweden showed that the amount of outsiders, the Conservatives own measure for how well the politics work, have increased with about 70 000 between 2006 and 2009. Yesterday, statistic showed that public assistance, the most serious kind of outsiderness, have increased in more than 90 percent of the municipalities last year.

Historically, the language use among the Social Democratic Party has not been as clear cut as in the case of the Conservative Moderate Party. Instead it seems that for the Social Democrats, the word 'outsiders' has been used in shifting contexts that gradually changed over time. Traditionally, they often tend to speak about 'outsiders' both in terms of unemployment, as well as, to a rather large extent, in relation to social 'outsiderness' and integration of citizens with their roots in foreign countries. However, from 2008 onwards, the Social Democrats tend to use the term in a similar way as the Conservative Moderate Party, i.e. mainly in relation to unemployment and labor-market issues. This is an interesting finding by itself to the extent that the Social Democrats, over time, have started to speak about 'outsiders' by the same terms as the Conservative Moderate Party. From a framing perspective, the Conservative Moderate Party has thus managed to make their language use on this specific issue to become the dominant frame for 'outsiders'.

---

19. http://www.socialdemokraterna.se/Vart-parti/Socialdemokratiska-riksdagsgruppen/Sonia-Karlsson/nyheter/Skapa-ett-nytt-trygghetssystem/

20. http://www.socialdemokraterna.se/media/nyheter/70-000-fler-i-utanforskap-med-moderaterna/

In an attempt to more systematically compare language use by the Social Democratic Party and the Conservative Moderate Party in relation to the 'outsider' concept, we have also performed a simple quantitative content analysis based on the documents from the parties' web pages.

From quality to quantity in party related documents

The national elections for the Swedish Parliament were held on September 19th, 2010. Table 1 shows the number of party documents containing the word 'outsiders' and the context of 'outsiders' between 2008 and 2010.

Table 1. Amount of party documents containing the word 'outsiders' and the context of 'outsiders' between 2008 to 2010 (percentages).

	Social Democratic Party				Conservative Moderate Party			
	2008	2009	2010	Total	2008	2009	2010	Total
outsider + unemployment	4	4	6	5	0	1	2	1
outsider + employment	9	9	7	8	8	6	9	8
outsider + work	17	19	17	17	23	27	26	26
outsider + jobs	23	22	21	21	18	15	26	20
outsider + subsidy	9	6	6	6	7	8	6	7
outsider + social	1	3	4	3	10	1	2	3
outsider + welfare	14	12	12	12	7	10	8	8
outsider + safety	9	10	9	9	8	8	10	9
outsider + insurance	3	4	2	2	0	2	0	1
outsider + segregation	1	1	3	2	0	4	1	2
outsider + integration	1	2	2	2	5	2	3	3
outsider + housing	3	2	3	3	0	2	2	2
outsider + povery	4	4	3	4	5	1	1	1
outsider + criminality	1	0	1	1	5	7	1	4
outsider + unsafety	3	0	3	2	3	2	1	2
outsider + abuse	0	1	1	1	2	3	1	2
outsider + insecurity	0	0	0	0	0	1	1	1
N: (contexts)	102	134	347	583	61	137	168	366
N: (documents)	34	39	122	195	19	61	68	148
Index of dispersion	.93	.92	.94	.94	.93	.92	.89	.91

Comment: The analysis is made as a simple word count for each year based on electronic documents available by the parties' web pages (www.socialdemokraterna.se and www.moderat.se). The percentages are based on the number of contexts. The contexts are in turn collected from the results of the qualitative content analysis of party documents. Index of dispersion is a simple measure for dispersion where a value of 1 indicates a totally even distribution of observations in each category while a value of 0 indicates the opposite. The measure is calculated as: $1 - \sum_{i=1}^{k} p^2$, where p is proportion while k stands for category.

Apparently, the number of documents containing the word 'outsiders' increased during 2010 for both parties, probably as a result of a growing number of documents in general during the election year. Clearly there is no sharp division in language use between the two parties. The word 'outsider' is mainly used in relation to employment issues and in relation to welfare and social security. In general, the Social Democratic Party is using the term 'outsiders' more frequently than are the Conservative Moderate Party. The language use in relation to the 'outsider' concept is, however, particularly stringent for both parties. When comparing the dispersion measures for the distribution in each category, the Conservative Moderate Party is a tiny bit less ambiguous in their language use during 2010 but the differences are still very small. To some sense, it seems as that the main difference is that the Social Democratic Party to some greater extent tend to speak about 'outsiders' in relation to the concepts of insurance and poverty.

Random Indexing of words related to 'outsider' in the Swedish blogosphere 2008–2010

In order to get an idea of how the word 'outsider' has been used in the Swedish blogosphere in the period from 2008 to 2010, we used RI to produce 1000-dimensional word spaces quarterly between November 2008 and September 2010.[21] For these experiments, we used word types as contexts in order to extract *semantic* relations from the word space, since we are interested in the *meaning* of the target term. Occurrence information was collected within a context window that spans two proceeding and two succeeding words, and ignores words with frequency less than 10. Table 2 shows the ten most semantically similar terms and their cosine similarities to 'outsider' for each quarter.

During the first four quarters from 2008 to late 2009 the 'outsider' concept seems to some extent to be related to housing situations by terms such as "overcrowding", "waiting lists" or "lack of housing". This is an interesting finding indeed since it, to a large extent reflects the way the Social Democratic Party was talking about 'outsiders' during earlier years between 2002 until 2008, that is in terms of segregation in the suburbs. In this respect, the results from our qualitative content analysis revealed an interesting pattern in that the language used by the Social Democrats during the last eight years subsequently adapted to the language used

21. The decision to conduct the studies on a quarterly year basis is more or less an arbitrary decision. With shorter time periods one risks to receive too many reference points with too little variation in, while longer time periods may contribute with variation but being less precise in discovering more sudden shifts in language use in the blogosphere.

Table 2. Words related to the term 'outsider' in the Swedish blogosphere between 2008 to 2010.

2008.Q4		**2009.Q1**		**2009.Q2**		**2009.Q3**	
Outsider:	***Prox.***	***Outsider:***	***Prox.***	***Outsider:***	***Prox.***	***Outsider:***	***Prox.***
fertile ground:	.25	long-term unemployed:	.34	subsidy dependency:	.39	poverty:	.33
insecurity:	.24	inscribed:	.34	contribution line:	.33	job-quake:	.31
exclusion:	.22	induced:	.22	segregation:	.30	admin. burdens:	.28
surplus:	.20	overcrowding:	.21	poverty:	.26	structural:	.26
safety net:	.19	taken cared of:	.20	unemployment:	.25	number of students:	.25
overcrowding:	.18	waiting lists:	.15	prosperity:	.24	range:	.23
network:	.18	contacted:	.15	lack of housing:	.23	lack of housing:	.23
conditions:	.18	poverty:	.15	labor shortage:	.23	segregation:	.23
fear:	.18	crime:	.14	homelessness:	.23	insecurity:	.21
criticized:	.18	predictions:	.14	insecurity:	.22	income inequalities:	.20
2009.Q4		**2010.Q1**		**2010.Q2**		**2010.Q3**	
Outsider:	***Prox.***	***Outsider:***	***Prox.***	***Outsider:***	***Prox.***	***Outsider:***	***Prox.***
job-quake:	.35	long-term unemployed:	.25	subsidy dependency:	.33	subsidy dependency:	.43
unemployment:	.30	alliance:	.25	need of sleep:	.31	poverty:	.30
subsidy dependency:	.28	unemployment:	.22	consumption space:	.28	need of sleep:	.24
health insurance:	.25	overlooked:	.21	welfare dependency:	.26	adjustment insurance:	.24
fewer:	.25	welfare:	.21	risk-taking:	.26	employment supply:	.24
poverty:	.22	segregation:	.20	poverty:	.26	insecurity:	.23
vulnerability:	.22	moderate-led:	.20	segregation:	.26	crime:	.23
security:	.22	fewer:	.20	needs:	.25	citizen participation:	.22
alliance:	.21	job-quake:	.20	influence:	.24	significantly:	.22
addiction:	.21	government:	.19	energy intake:	.24	community policing:	.22

Comment. The table shows the top-ten neighboring words to the concept of 'outsiders'. The division of the time-periods is done quarterly from October 2008 to September 2010. The proximity measures are the cosine angles between the context vectors. It should be noted that every time an experiment is run using RI, the resulting word space will be slightly different, partly due to the impact of parameter-settings and partly due to the very fact that we are using random indexing as a base for the experiment. This means that (1) that the similarity measures are highly relative (and hence only informative in relation to other similarity measures in the same model) and (2) that there may be some variance in the analysis. One way to overcome these issues is to run the simulations multiple times with different random seeds and different parameters, and only report the robust results. The original Swedish terms are listed in the Appendix.

by the Conservative Moderate Party when speaking about 'outsiders'. The result from the RI analysis of the earlier time-periods of the blogosphere could thus be an indication that the language use in the blogosphere is to some extent lagging behind in relation to the language used by the parties in general and for the Social Democratic Party in particular, given that the latter are the agenda setters in this respect. The concepts relating to housing situations is, however, not used to any greater extent in relation to the 'outsider' concept by any of the parties during the investigated period for the RI analysis in any of neither documents.

Another indication of the impact of the Social Democratic language use during 2008 and early 2009 is the occurrence of the term 'poverty', which is one of the most closely related terms to the 'outsider' concept (Table 2). In the previous analysis of the party documents there were some indications that the concepts of 'poverty' and 'insecurity' to some extent were more frequently used in relation to 'outsiders' by the Social Democrats. The presence of terms relating 'outsiders' to concepts of housing situations, poverty and segregation, etc., should, however, not be exaggerated since the most widely related concepts by far pertain to 'subsidy dependency' and 'long-term unemployment', which are concepts that are mainly used by both parties in relation to 'outsiders' over this period.

However, especially from late 2009 onwards, words such as 'job-quake', 'contribution line', 'subsidy dependency' and 'Alliance' score higher in the word space. Interestingly, these terms are directly related to concepts invented and used by the Conservative Moderate Party and its alliance parties. At the same time, words such as 'poverty' and 'segregation' are scoring lower. This could be an indication that the language use by the Conservative Moderate Party and its alliances are becoming the agenda-setting language when it comes to defining the 'outsider' concept before the election campaign.

Nevertheless, based on the results from the quantitative content analysis of the party documents, we cannot discern any sharp divisions in the language use in relation to the 'outsider' concept by any of the two parties during the investigated time period from late 2008 onwards. There are some indications that the Social Democrats to some extent are a bit more ambiguous in their language use. For example, they are, as mentioned, using the 'outsider' concept in relation to poverty and insecurity. However, in general both parties tend to speak about 'outsiders' in relation to the labor market. A striking difference is, however, that while the Social Democratic Party tends to speak about employment and unemployment the Conservative Moderate Party tends to speak more one-sidedly of employment. Both these terms are also apparent in the RI word space. Nevertheless, given the results from prior analysis of the party related documents it is difficult, not to say impossible, to validate which parties' language use is dominating the blogosphere during the investigated time period.

Table 3. Proximity measures for the concept of 'outsiders' and the Conservative Moderate Party and the Social Democratic Party in the blogosphere between 2008 and 2010.

2008.Q4	***Freq.***	***Prox.***	**2009.Q1**	***Freq.***	***Prox.***
moderate + outsider	122 (/3961)	0.11	moderate + outsider	95 (/4019)	0.11
social democrat + outsider	74 (/4141)	.08 (.11)	social democrat + outsider	72 (/4856)	.09 (.01)
2009.Q2			**2009.Q3**		
moderate + outsider	149 (/7134)	0.06	moderate + outsider	274 (/6408)	0.16
social democrat + outsider	184 (/8365)	.11 (.10)	social democrat + outsider	286 (/7255)	.11 (.10)
2009.Q4			**2010.Q1**		
moderate + outsider	311 (/5859)	.20	moderate + outsider	300 (/7158)	.20
social democrat + outsider	300 (78365)	.18 (.16)	social democrat + outsider	263 (/6888)	.17 (.13)
2010.Q2			**2010.Q3**		
moderate + outsider	145 (/17849)	.13	moderate + outsider	99 (/25194)	.10
social democrat + outsider	85 (/19966)	.10 (.10)	social democrat + outsider	77 (/25266)	.12.(10)

Comment. In this table we used documents, i.e. single blog posts as contexts in order to compute *associative* relations from the word space. The table shows the total amount of blog posts that contains the words "moderat" or "social democrat" in parentheses, and the amount of blog entries that contains both the words "moderat" or "social democrat" + "outsider". The proximity measures (which is the cosine angles between the context vectors) within parentheses are for the slang expression "sossar", which sometimes is used instead of "Social Democrats".

In an attempt to get a more systematic comparison of the connection between the terms related to the 'outsider' concept in the blogosphere and the political parties, we counted the amount of blog entries where the party labels and the word 'outsider' are mentioned simultaneously. In connection to this saliency analysis we also conducted an RI-analysis for the same time period. In this experiment, we used documents (i.e., blog posts) as contexts in order to compute associative relations, since we are interested in the extent to which the term 'outsider' is associated with the two major parties. The results from this part of the analysis can be found in Table 3.

Table 3 shows the number of blogs that contain the words 'Moderate' or 'Social Democrat' together with the word 'outsider'. Considering the particularly low proximity values, it is unclear if one can really draw any conclusions (since such low similarity measure indicates that there is a lot of noise in the context vectors). For this reason it may be better to simply compare the frequencies of the word 'outsider' in blog entries also containing the words 'Moderate' and 'Social Democrat'.

In general, the same number of blog entries occurs that contains the words 'Moderate' and 'Social Democrat', but the term 'outsider' seems to occur more frequently in messages containing the word 'Moderate'. As mentioned above, the results from the previous word space reveal that more specific terms were directly related to concepts invented and used by the Conservative Moderate Party and its alliance parties. Terms such as 'job-quake', 'contribution line', 'subsidy dependency', tended to be more dominant in the blogosphere from late 2009 to mid 2010. A cautious interpretation against this background would thus be that these results together give some support to conclude that the language use of the Conservative Moderate Party tends to be a bit more dominant in the blogosphere when it comes to the concept of 'outsiders'. In other words, the Conservative Moderate Party seems to have been those that have set the scene for how one should define and understand the concept of 'outsiders'.

Summary and conclusions

The aim of this study has been to compare the language use in relation to the targeted concept of 'outsiders' in party related documents with the language use in the Swedish blogosphere. The analysis of the party related documents from 2008 and onwards did, however, indicate a highly similar language use among the Social Democratic- and the Conservative Moderate Party concerning the concept of 'outsiders'. The implication for our study in this respect was that we were not able to identify any specific agenda setter in terms of language use, due to the limitations in blogosphere data over time. Nevertheless, the results from the RI

analysis have validity. The top ten semantic similarities from the RI word space model of the Swedish blogosphere were all well-known terms and concepts that could be derived from the party related documents. From this perspective, we have after all been able to identify how the term 'outsider' has been framed by the political parties and, hence, how this has contaminated the language use in the blogosphere during late 2008 to 2010.

For example, in the beginning of the period from late 2008 to early 2009, we were able to identify words related to housing situations and segregation in connection to the 'outsider' concept in the Swedish blogosphere. Talking about 'outsiders' in these terms was something that the Social Democratic Party tended to do in the years before 2008 (the language use in the blogosphere seems to be lagging behind in relation to the language used by the Social Democratic Party during the investigated period). In the middle of the investigated time span, in the period from late 2009 to mid 2010, the results from the RI word space revealed specific terms in relation to 'outsiderness' that were directly connected to concepts invented and used by the Conservative Moderate Party and its alliance parties. We also found that the term 'outsider' tended to occur more frequently together with the word 'Moderate' than for 'Social Democrat' in different blog entries. A cautious interpretation against this background would thus be that the results all together give some support for that the language use by the Conservative Moderate Party tends to be a bit more indicative in the Swedish blogosphere when it comes to the concept of 'outsiders', especially since the Social Democratic Party historically has a tendency to adapt their language use in relation to the 'outsider' concept to the language used by the Conservative Moderate Party.

References

Chong, D. and N. J. Druckman. 2007. Framing public opinion in competitive democracies. *American Political Science Review* 101(4), pp. 637–655. DOI: 10.1017/S0003055407070554

Kanerva, P., J. Kristofersson and A. Holst. 2000. Random Indexing of Text Samples for Latent Semantic Analysis. *Paper presented at the Proceedings of the 22:nd Annual Conference of the Cognitive Science Society*.

Kinder, D. R. 2003. Communication and politics in the age of information. In D. O. Sears, L. Huddy and R. Jervis, (eds.), *Handbook of Political Psychology*. Oxford: Oxford University Press.

Naurin, D., R. Eising, C. Mahoney, D. Coen, F. Baumgartner and S. Sauragger. 2009. A Systematic Comparative Research Project on Interest Group Politics in Europe. *European Collaborative Research Projects in the Social Sciences (ECRP)*.

Oscarsson, H. and S. Holmberg. 2008. *Regeringsskifte*. Stockholm: Norstedts juridik.

Rueda, D. 2005. Insider-outsider politics in industrialized democracies: the challenge to social democratic parties. *American Political Science Review* 99(1), pp. 61–74. DOI: 10.1017/S000305540505149X.

Rueda, D. 2006. Social democracy and active labour-market policies: insiders, outsiders and the politics of employment promotion. *British Journal of Political Science* 36(3), pp. 385–406. DOI: 10.1017/S0007123406000214

Sahlgren, M. 2006. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. Stockholm: Stockholm University.

Sahlgren, M. 2005. An introduction to random indexing. *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*, TKE 2005, August 16, Copenhagen, Denmark.

Schütze, H. 1993. Word space. In *Proceedings of the 1993 Conference on Advances in Neural Information Processing Systems*. San Francisco: Morgan Kaufmann Publishers Inc., pp. 895–902.

Slothuus, R. 2008. *How Political Elites Influence Public Opinion: Psychological and Contextual Conditions of Framing Effects*. Århus: Politica.

Slothuus, R. and C. de Vreese. (2008). Political parties, motivated reasoning, and issue framing effects. Unpublished manuscript, Amsterdam.

Turney, P. and P. Pantel. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research (JAIR)* 37(1), pp. 141–188. AI Access Foundation.

Walgrave, S. and K. De Swert. 2007. Where does issue ownership come from? From the party or from the media? Issue-party identifications in Belgium, 1991–2005. *The International Journal of Press/Politics* 12(1), pp. 37–67.

Appendix

(1) Målet med förändringen av systemet är att hjälpa människor tillbaka i arbete och minska gapet mellan utanförskap och arbete.

(2) Vi väljer att slå vakt om Sveriges sunda offentliga finanser. Vi väljer att föra en tydlig politik för full sysselsättning och minskat utanförskap.

(3) Under Socialdemokraternas tid i regeringen ökade sjukskrivningarna kraftigt. För att förbättra statistiken förde man över 140 personer om dagen till förtidspension och ett permanent utanförskap.

(4) Det nödvändigt att se vad som ligger bakom brottsligheten: ofta rör det sig om brist på jobb, brist på sociala gemenskaper och brist på trygghet. Därför behöver vi arbeta ännu mer för att bryta utanförskapet: för fler i arbete, för nya arbetstillfällen, för en bättre skola och för att nyanlända ska integreras. På så sätt får fler barn en trygg uppväxt – den mest effektiva brottsförebyggande insatsen. Genom att arbeta mot utanförskap arbetar vi också mot brottsligheten, detta är ett samband så starkt att det ibland glöms bort.

(5) Brottsligheten bekämpas genom ett starkt rättsväsende och genom en politik som minskar utanförskap och arbetslöshet.

(6) Ett stort antal i utanförskap hotar välfärden eftersom färre bidra till att skapa de begränsade resurser som måste delas av fler människor. Regeringens ansträngningar för hög sysselsättning är det enda långsiktigt hållbara sättet att skydda välfärden. (---) De åtgärder som har vidtagits för att minska utanförskapet bidrar också till att säkra pensionerna.

(7) Vad är svaret från regeringen Reinfeldt idag? Utanförskapet ökade i juni för tredje månaden i rad.
(---) Det har gått en halv mandatperiod. Det är dags för utvärderingen. Det är upp till bevis för regeringen.
Gjorde det vad de sa att de skulle göra? Hur gick det med jobben? Hur gick det med utanförskapet? Hur gick det med välfärden och tryggheten? Vad är regeringen Reinfeldts besked idag?

(8) Sysselsättningen behöver öka – bland unga, bland invandrade svenskar, bland deltidsarbetande kvinnor och, inte minst, bland äldre. Ska vi nå dit måste den ofrihet som stavas arbetslöshet och utanförskap bekämpas – och för den delen ofrivillig deltidsarbetslöshet. Ska vi nå dit måste ”arbetslinjen” gälla alla.

(9) Under det senaste året har antalet arbetslösa ökat med 100 000. Sedan valet 2006 har utanförskapet vuxit med 70 000 människor,. (---) I valet 2010 kommer arbetslösheten att vara den allt annat överskuggande frågan.

(10) För snart fyra år sedan vann de borgerliga partierna valet på att utlova fler jobb och färre i utanförskap. När vi nu kan utvärdera resultatet av mandatperioden är det dessvärre nedslående. Regeringen kastar ut tiotusentals svenska medborgare ur sjukförsäkringen in i försäkringslöshet. På bara ett år har antalet arbetslösa ökat med fler än 100 000 personer, och antalet i utanförskap med 70 000 personer. Utbetalningarna av socialbidrag beräknas öka med 50 procent mellan 2006 och 2011.

(11) Idag kom siffror från SCB som visar att utanförskapet, Moderaternas eget mått på hur väl politiken fungerar, har ökat med ca 70 000 mellan 2006 och 2009. Igår kom siffror som visar att socialbidragen, den allvarligaste formen av utanförskap, ökat i över 90 procent av kommunerna förra året.

Table 2. Words related to the term 'outsider' in the Swedish blogosphere between 2008 to 2010 (in Swedish).

2008.Q4		**2009.Q1**		**2009.Q2**		**2009.Q3**	
Utanförskap:	***Prox.***	***Utanförskap:***	***Prox.***	***Utanförskap:***	***Prox.***	***Utanförskap:***	***Prox.***
grogrund:	0.25	långtidsarbetslös:	0.34	bidragsberoende:	0.39	fattigdom:	0.33
otrygghet:	0.24	inskriven:	0.34	bidragslinjen:	0.33	jobbävning:	0.31
utanförskapet:	0.22	förmåtts:	0.22	segregation:	0.30	regelkrångel:	0.28
mervärden:	0.20	trångboddhet:	0.21	fattigdom:	0.26	strukturell:	0.26
skyddsnät:	0.19	omhändertagit:	0.20	arbetslöshet:	0.25	elevantal:	0.25
trångboddhet:	0.18	bostadsköer:	0.15	välstånd:	0.24	omfång:	0.23
nätverk:	0.18	kontaktats:	0.15	bostadsbrist:	0.23	bostadsbrist:	0.23
förutsättningar:	0.18	fattigdom:	0.15	arbetskraftsbrist:	0.23	segregation:	0.23
rädsla:	0.18	kriminalitet:	0.14	bostadslöshet:	0.23	otrygghet:	0.21
förkättrade:	0.18	förutsägelser:	0.14	otrygghet:	0.22	inkomstklyftor:	0.20
2009.Q4		**2010.Q1**		**2010.Q2**		**2010.Q3**	
Utanförskap:	***Prox.***	***Utanförskap:***	***Prox.***	***Utanförskap:***	***Prox.***	***Uanförskap:***	***Prox.***
jobbävning:	0.35	långtidsarbetslöshet:	0.25	bidragsberoende:	0.33	bidragsberoende:	0.43
arbetslöshet:	0.30	alliansen:	0.25	sömnbehov:	0.31	fattigdom:	0.30
bidragsberoende:	0.28	arbetslöshet:	0.22	konsumtionsutrymme:	0.28	sömnbehov:	0.24
sjukförsäkring:	0.25	skymundan:	0.21	socialbidragsberoende:	0.26	omställningsförsäkring:	0.24
färre:	0.25	välfärd:	0.21	risktagande:	0.26	arbetsutbud:	0.24
fattigdom:	0.22	segregation:	0.20	fattigdom:	0.26	otrygghet:	0.23
utsatthet:	0.22	moderatledda:	0.20	segregation:	0.26	kriminalitet:	0.23
regelverk:	0.22	färre:	0.20	bolag:	0.25	medborgarinflytande:	0.22
alliansregeringen:	0.21	socialbidragsberoende:	0.19	anmälningsplikt:	0.24	klimatmål:	0.22
missbruk:	0.21	regeringen:	0.19	energiintag:	0.24	närpoliser:	0.22

CHAPTER 5

Text to ideology or text to party status?*

Graeme Hirst, Yaroslav Riabinin, Jory Graham,
Magali Boizot-Roche, and Colin Morris
Department of Computer Science, University of Toronto

Several recent papers have used support-vector machines with word features to classify political texts – in particular, legislative speech – by ideology. Our own work on this topic led us to hypothesize that such classifiers are sensitive not to expressions of ideology but rather to expressions of attack and defence, opposition and government. We tested this hypothesis by training on one parliament and testing on another in which party roles have been interchanged, and we find that the performance of the classifier completely disintegrates. But removing the factor of government–opposition status, as in the European Parliament, enables a more-ideological classification. Our results suggest that the language of attack and defence, of government and opposition, may dominate and confound any sensitivity to ideology in these kinds of classifiers.

1. Introduction

There have been a number of attempts recently to develop methods to automatically determine the ideological position of a political text. For example, one might wish to take a newspaper editorial or a blog and classify it as socialist, conservative, or Green. In practice, much of the research has taken speeches by members of a legislature (such as the U.S. Congress or the European Parliament) as the text to be classified and indicators such as party membership or legislative voting patterns as a proxy for ideology (indeed, Yu et al. (2008) use the terms *party classifier* and *ideology classifier* almost interchangeably); thus the problem becomes one of predicting one of these indicators from speech. One might expect, a priori, that

* This is an extended version of "Party status as a confound in the automatic classification of political text" by Graeme Hirst, Yaroslav Riabinin, and Jory Graham, *Proceedings, 10th International Conference on Statistical Analysis of Textual Data (JADT 2010),* Rome, June 2010.

This work is financially supported by the Natural Sciences and Engineering Research Council of Canada.

methods based solely on the vocabulary used in a text would not be effective, because the members of a legislature, regardless of ideology, are all discussing the same topics – e.g., the legislation before them or the issues of the day – and hence would all be using the same topic-derived vocabulary (Mullen and Malouf 2006). The ideology expressed in a text would thus be apparent only at the sentence- and text-meaning levels. Nonetheless, one might hypothesize that different ideological frameworks lead to sufficiently different ways of talking about a topic that vocabulary can be a discriminating feature (Lin et al. 2006). And indeed, several studies have obtained notable results merely from classification by support-vector machines (SVMs) with words as features ('bag-of-words classification').[1]

For example, Thomas et al. (2006) examined speeches made by members of the U.S. House of Representatives to try to determine whether each speaker supported or opposed the proposed legislation under discussion. They combined bag-of-words text classification by SVMs with textual information about each speaker's agreement or disagreement with other speakers, obtaining an accuracy of around 70% (the majority baseline was 58%). Greene (2007) obtained an improved accuracy of over 74% on the same task by annotating each word with its grammatical relation from a dependency parse. Jiang and Argamon (2008), on the related task of classifying political blogs as liberal or conservative, improved results over using word features of the whole text by first trying to identify subjective sentences and the expressions of opinion that they contain, and then limiting the features to those parts of the text.

Diermeier et al. (2007) used SVMs with bag-of-words features to classify members of the U.S. Senate by ideology, labelling each speaker as a liberal or a conservative, and achieved up to 94% accuracy. However, in these experiments, the authors focused on 'extreme' senators – the 25 most conservative and the 25 most liberal members in each Senate. On 'moderate' senators, the results were notably poorer (as low as 52% accuracy). Moreover, there was considerable overlap between the training and testing portions of Diermeier et al.'s dataset, since they extracted content from multiple Senates (101st to 108th) and since members of Congress tend to preserve their beliefs over time. Specifically, 44 of the 50 'extreme' Senators in their test set were also represented in the training data, which means that the classifier was already trained on speeches made by these particular individuals. Thus the classifier might be learning to discern speaking styles rather than ideological perspectives.

1. Observe that this goal differs from that of, e.g., Gryc and Moilanen (this volume) and Dahlberg and Salhgren (this volume), who aim to determine the position expressed in a text with regard to a particular topic, such as Barack Obama or 'outsiders' in Sweden. By contrast, the more general goal here is the ideological position underlying a text, independent of any particular topic.

Later work by the same authors (Yu et al. 2008) made no distinction between moderates and extremes; rather, they tried to classify all members of the 2005 U.S. Congress by party affiliation, achieving an accuracy of 80.1% on the House of Representatives and 86.0% on the Senate. The goal of their study was to examine the person- and time-dependency of the classifier by using speeches from both the Senate and the House and comparing the results. They found that party classifiers trained on House speeches could be generalized to Senate speeches of the same year, but not vice versa. They also observed that classifiers trained on House speeches performed better on Senate speeches from recent years than older ones, which indicates the classifiers' time-dependency.

We began the present work to see whether these kinds of bag-of-words SVM classification methods would hold up in analysis of speech in the Canadian Parliament (Section 3 below). Our results, however, led us to question whether vocabulary differences between parties really reflected ideology or whether they had more to do with each party's role in the Parliament, and we investigate this in Section 4 below.

2. Background: The Canadian party system and Parliament

The Canadian Parliament is a Westminster-style parliament. The party with the most seats in the House of Commons (albeit possibly a minority of them) forms the government; the other parties are the opposition. There may also be a few Independent (unaffiliated) members. In the last 12 years, there have been four or five parties in each Parliament. In broad terms the parties may be classified as conservative (Reform Party, Canadian Reform Conservative Alliance, Progressive Conservative Party, Conservative Party of Canada), liberal or centre (Liberal Party),[2] or left-wing (New Democratic Party and Bloc Québécois); see Collette and Pétry (this volume) for more discussion of the parties' left–right positions.

Both English and French are official languages of Canada. A speaker in Parliament may use either language, and will sometimes even switch between the two within a speech. Everything said in Parliament is professionally translated into the other official language, and the proceedings are published in both languages. Thus the published English text of the debates is a mixture of original English and translations from French, and the French text has the complementary distribution.

2. Thus in our data, all liberals are Liberals, but not all conservatives are Conservatives. Similarly, we distinguish between opposition parties – any party that is not the governing party – and the Opposition party – the opposition party from which the Leader of the Opposition is drawn.

3. First set of experiments: Classifying by party

The present work was intended as a prelude to a larger project on ideological analysis of text. Our first task, intended as a baseline, was to apply bag-of-words support-vector machine classification, as used by Diermeier et al. (2007) and Yu et al. (2008) on U.S. Congressional speech, to speech in the Canadian Parliament, to see whether we could classify the speech by party affiliation (as a proxy for ideology) and obtain similar results, despite the differences in the political systems of the two countries.

In Canadian politics, unlike those of the U.S., party discipline is strong and (with only rare exceptions) all members of a party will vote the same way. The governing party will always vote to support its legislation; an opposition party might oppose it or support it. Thus (in contrast to the tasks described by Diermeier et al. and Yu et al.), there is no meaningful distinction between predicting voting records from parliamentary speech and predicting party affiliation. On one hand, it might be argued that this makes the task easier because parliamentary speech is likely to be highly partisan. On the other hand, it might be argued that it makes the task more difficult, because there is a greater diversity of views with precisely the same voting pattern, and so the classification is less straightforward.

In order to avoid the problems inherent in cross-time analysis, as highlighted by the work of Diermeier et al. (2007), we focus in this section on a single time period, so that there is a one-to-one mapping between members of Parliament (MPs) and documents in our dataset. Each document is a concatenation of all the speeches made by a speaker, and no other document contains text spoken by that person. Thus no speaker appears in both training and test data.

3.1 Data

We used both the English and French *House of Commons Debates* ('Hansard') for the first 350 sitting days of the 36th Parliament (1997-09-22 to 2000-05-10). In the 36th Parliament, a majority government was formed by the Liberal Party, led by Jean Chrétien. This data was available in a convenient plain-text form with sentence breaks identified (Germann 2001), as it has been widely used for research in machine translation.

We considered two sections of the proceedings: the debates on legislation and other statements by members ('Government Orders') and the oral question period. And we focused on the governing Liberal Party and the opposition conservative

parties,[3] in order to do a binary discrimination, liberal versus conservative; the left-wing parties had relatively few members in this Parliament and were excluded from the analysis.

For each MP who was a member of one of the liberal or conservative parties, and for each language, we formed a 'document' by concatenating all their utterances in debates, question period, or both, throughout the Parliament. (For simplicity, we will refer to all utterances as 'speeches', regardless of their length, including questions and answers in the oral question period.) We experimented with a variety of pre-processing methods, including stemming the words or leaving them whole, removing or retaining stopwords (defined as the 500 most frequent words in the text), and removing or retaining rare words (defined as those occurring in fewer than five documents). (Details of these and other pre-processing matters are given by Riabinin 2009.) In some of our experiments, we discarded the data for members who said very little, or nothing at all, in question period or in debates, using 200 documents representing 121 liberals and 79 conservatives; in other experiments, we considered all 156 liberal and 79 conservative members who spoke at all.[4] In all, depending on our choices in pre-processing, we had about 4 million words in each language for liberals (of which approximately 900,000 were from the question periods) and 2.7 million for conservatives (of which approximately 500,000 were from the question periods).

Generally, these variations in pre-processing made little difference to the results. In this paper we report results for experiments on the texts for all speakers, with words left unstemmed and with rare words removed, which usually, though not invariably, gave the best results.

3.2 Method

Taking word-types as the features for classification – that is, regarding the document for each speaker as a bag of words – for each language we trained an SVM classifier for ideology as indicated by party membership, liberal or conservative.

3. At the time of this Parliament, the conservative parties were in disarray. The Opposition was the conservative Reform Party (which became the Canadian Reform Conservative Alliance in March 2000), but the conservative Progressive Conservative Party also held a number of seats.

4. Several members of the conservative parties either defected to the Liberal party or became independents during this Parliament; and one member of the left-wing NDP defected to a conservative party. We treated all these members as conservatives in our experiments; for details and rationale, see Riabinin (2009).

In training and testing, we used five-fold cross-validation. We experimented with four weighting schemes: *boolean* (presence of feature), *tf* (term frequency), *tf-norm* (term frequency normalized by document length), and *tf-idf* (term frequency by inverse document frequency). The best results were obtained with *tf-norm* and *tf-idf*; the results we present below all use the latter.

3.3 Results

Table 1 shows the accuracy of classification of party membership by the SVM for each language on the documents of each data set: oral question period (OQP), debates (GOV), and the two combined (OQP + GOV). In all cases, retaining the 500 most frequent features led to higher accuracy than removing them. The baseline method of choosing the larger class (liberal) for all members would give an accuracy of 65.5%. All our results are well above this baseline, and in fact reach almost 97% for oral question period in English when frequent words are retained. The reason for the discrepancy between this result and the 89.5% obtained for the same data in French is unclear, as the two texts are mutual translations and no such effect was seen with the debates texts.[5] We also observe that in three cases out of four, combining debates and question period in a single classifier is deleterious to accuracy compared to classifying each separately. Generally speaking, our results are similar to, or better than, those of Yu et al. (2008) on the U.S. Congress.

Table 1. Accuracy (%) of classification by ideology on speech in the oral question period (OQP) and debates (GOV) by liberal and conservative members of the 36th Parliament, with and without removal of the 500 most frequent features (majority baseline = 65.5%).

	OQP + GOV	OQP	GOV
With 500 most frequent features retained			
English	83.8	96.9	83.3
French	83.2	89.5	86.0
With 500 most frequent features removed			
English	78.7	92.9	79.6
French	80.8	84.8	83.5

5. Compare the results of Collette and Pétry (this volume) on the differences that they found in locating English and French political manifestos on a left–right spectrum, and the differences between languages that they adduce in explanation. In our results, however, while the accuracy obtained for each language sometimes varies quite noticeably within each condition, there is no apparent system in the differences; sometimes the English results are more accurate and sometimes the French results are; sometimes the difference is marginal.

3.4 Discussion

The higher accuracy of classification for question period than for debates suggests that the language of question period is in some way more partisan than that of debates. However, our examination of the most discriminative words suggests that this partisanship is not so much ideological as a matter of attack and defence. In particular, in the Canadian Parliament, the oral question period consists largely of hostile questions from members of the opposition parties to ministers of the government, with only occasional friendly questions from government backbenchers, which themselves often serve primarily to set up an attack on the opposition.[6] It's possible, therefore, that our classifier may be learning – at least in part – not to distinguish ideologies but to distinguish questions from answers or attack from defence, which is not the goal of our research. Table 2 shows the ten most discriminative English words for each side in question period. For the governing liberals, the top words are *hon* and *member*, as in *the hon. member for Halifax West*, which is how a minister from the governing party typically addresses a member who has asked a question. Also, the word *we* might be used by a minister to speak on behalf of the entire party or government when responding to questions. For the opposition conservatives, the word *why* serves the obvious purpose of posing a question, and the words *he* and *her* are likely used to refer to government ministers who are the targets of the questioning. Also, observe the use of words such as *bloc*, *reform*, and *opposite* by the liberals, and *prime* (as in *Prime Minister*) and *liberal* and *liberals* by the conservatives.[7] This lends further support to the hypothesis that the classifier is partially learning to distinguish government members from opposition members.

When frequent words are removed we see this effect less, with a corresponding drop in accuracy (see the second part of Table 1), but it does not disappear entirely. In this condition, we certainly see reflections of ideology in vocabulary. The liberal lexicon is characterized by words related to Québec (*French*, *Francophonie*, *MAI* [*Montréal Arts Interculturels*], *PQ* [*Parti Québécois*]) and various social issues (*housing, violence, humanitarian, youth, society, technology*), while the conservatives tend to focus on monetary concerns (*APEC, taxpayer, dollar, millions, paying, premiums*), aboriginal affairs (*native, Indian, chief*), and, to a lesser degree, national defence (*military, marshall*). Nonetheless, the governing liberals use language that

6. This contrasts with the practice in similar parliaments, such as those of Australia and the U.K., in which questions are more evenly balanced between those of the opposition and those of government backbenchers.

7. Interestingly, this tendency for the names of opponents to be discriminating features is the converse of what Lin et al. (2006) found in their analysis of an Israeli–Palestinian debate, in which naming one's own side was discriminating, but see Section 6.4.2 below.

Table 2. The top 10 English words characterizing each class in the oral question period.

Rank	liberal (government)	conservative (opposition)
1	hon	prime
2	member	why
3	we	liberal
4	opposite	solicitor
5	quebec	farmers
6	housing	finance
7	bloc	he
8	reform	liberals
9	québécois	hrdc[a]
10	women	banks

[a] HRDC = Human Resources Development Canada, a federal government department.

is generally positive (*congratulate, excellent, progress*) and is intended to create the appearance of a government at work (*established, inform, improve, assist, developing, promote*). In contrast, the opposition conservatives use negative words that are meant to call the government's competence into question (*justify, resign, failed, admit, refusing, mismanage*). So again, it seems that many of the features relate not to ideology but to attack and defence – not to the party's beliefs but to its status as government or opposition.

4. Second set of experiments: Classifying by party status

Even if a classifier for political speech were truly using features related to ideology, we would expect that at least some of these features would specifically pertain to views of current events and therefore, if it is trained on one Parliament, it will not perform as well on a different Parliament in which different events are current, as in the results of Diermeier et al. discussed in Section 1 above. Nonetheless, we would expect that many of the features will be invariant over time and that such a classifier will still perform much better than a baseline.

On the other hand, if the 'ideological' classifier is in reality using (solely or primarily) features related to government and opposition status, then training on one Parliament would carry over only to other Parliaments in which the parties hold the same status; if they swap roles, then the classifier will fail. Indeed, in such a case it might (or should!) perform *worse* than the majority baseline, tending to classify liberals as conservatives and vice versa. In our second set of experiments, we tested the hypothesis that the latter is the case that an SVM bag-of-words

classifier for Canadian parliamentary speech is primarily sensitive to party status, not ideology. We also looked at the in-between case: training an 'ideological' classifier on data in which all combinations of ideology and party status are present.

4.1 Data

To test our hypothesis, we needed a Parliament in which, in contrast to the 36th Parliament, a conservative party was in government. We chose the recent 39th Parliament (2006-04-03 to 2008-09-07), with a minority Conservative Party[8,9] government led by Stephen Harper; the Liberal Party was in opposition, along with the New Democratic Party and the Bloc Québécois. The proceedings were downloaded from the Parliament of Canada website in HTML-formatted documents and processed into a format similar to that of the 36th Parliament data.

4.2 Method and results

4.2.1 *Replication of the first experiments on the new data*

We first replicated the experiments of Section 3 on the new data, discriminating liberal members from conservative members (there was sufficient data for 104 liberals and 130 conservatives) within the same Parliament. Training and testing with five-fold cross-validation on the 39th Parliament, we achieved results similar to those of the 36th Parliament, albeit with slightly lower accuracy, especially for the English OQP documents; see Table 3 and compare Table 1. In particular, the accuracy of the classification on French text of speakers in Government Orders is anomalously low (baseline level) compared to all our other results including those for the English translation of the same text; we have no explanation for this. We also observe that for this data, unlike the 36th Parliament, the strategy of removing the 500 most frequent words is sometimes superior to that of retaining them.

Examining the primary features used in the classification for oral question periods, we observed that several words 'swapped sides': four of the top 10 English words that characterized the liberals in the 36th Parliament characterized conservatives in the 39th Parliament, and the primary word that characterized conservatives in the 36th Parliament was the second word that characterized liberals in the 39th; see Table 4. This is evidence for our hypothesis that the classifier is really picking up features related to government and opposition status.

8. So in this Parliament, unlike the 36th, all conservatives are Conservatives.

9. http://www2.parl.gc.ca/housechamberbusiness/ChamberSittings.aspx

Table 3. Accuracy (%) of classification by ideology on the 39th Parliament, with and without the 500 most frequent words retained (majority baseline = 55.8%).

	OQP + GOV	OQP	GOV
With 500 most frequent features retained			
English	83.8	88.3	72.3
French	75.5	88.8	56.8
With 500 most frequent features removed			
English	79.9	83.5	73.2
French	79.0	88.2	57.2

Table 4. The top 10 English words characterizing each class in oral question periods in each Parliament (extending Table 2). Boldface indicates words that 'swap sides' between the two Parliaments. Boldface italic words characterize the governing side; the boldface roman word characterizes the opposition.

Rank	36th Parliament		39th Parliament	
	liberal (government)	conservative (opposition)	liberal (opposition)	conservative (government)
1	hon	**prime**	conservatives	***bloc***
2	***member***	why	**prime**	liberals
3	***we***	liberal	conservative	senate
4	opposite	solicitor	immigration	violent
5	quebec	farmers	mulroney	***we***
6	housing	finance	kyoto	***québécois***
7	***bloc***	he	admit	greenhouse
8	reform	liberals	minority	ndp
9	***québécois***	hrdc	promise	corruption
10	women	banks	her	***member***

4.2.2 *Classifying across Parliaments*

Again we used the proceedings of the 36th and 39th Parliaments, both English and French, but in each language we took the classifiers trained on one Parliament and tested them on the other. (In these experiments, we have the deprecated situation that some individual speakers, being members of both parliaments, occur in both the training data and the test data and thereby might give the classifier an unfair boost.) The results, shown in Table 5, are in all cases well below the majority baseline scores, just as we hypothesized; when party status changes, there are no constant ideological features to save the classifier.

We also tried training classifiers on the data of the two Parliaments combined. This dataset includes all combinations of ideology and party status – that is liberals in government, liberals in opposition, conservatives in government,

Table 5. Accuracy (%) of classification by ideology when training on one Parliament (36th or 39th) and testing on the other.

Training → Testing	**OQP + GOV**	**OQP**	**GOV**
36 → 39 *(Majority baseline = 55.8%)*			
With 500 most frequent features retained			
English	44.9	43.3	44.6
French	45.7	46.1	47.0
With 500 most frequent features removed			
English	46.2	44.6	44.1
French	43.5	49.6	43.5
39 → 36 *(Majority baseline = 65.5%)*			
With 500 most frequent features retained			
English	36.8	34.5	36.2
French	35.2	51.1	33.5
With 500 most frequent features removed			
English	35.0	49.6	42.7
French	36.4	51.1	33.5

and conservatives in opposition. Some speakers, those who were members of both Parliaments, appear with each possible party status, whereas others, those who were members of only one of the two Parliaments, appear in only one of these four conditions. A classifier trained on the former group performs at around the level of the majority baseline (Table 6); one trained on the latter does better (Table 7), but the results are overall below the level of the original experiments (Tables 1 and 3), especially for OQP data. (The exception is that the anomalously low results for French GOV data are not seen when frequent features are retained.)

Table 6. Accuracy (%) of classification by ideology on speakers who were members of both the 36th (liberal government) and 39th Parliament (conservative government), with and without the 500 most frequent words retained (majority baseline = 64.0%).

	OQP + GOV	**OQP**	**GOV**
With 500 most frequent features retained			
English	62.0	66.9	61.1
French	63.0	63.0	63.0
With 500 most frequent features removed			
English	64.0	66.9	59.4
French	64.0	64.0	64.0

Table 7. Accuracy (%) of classification by ideology on speakers who were members of either the 36th (liberal government) or 39th Parliament (conservative government), but not both, with and without the 500 most frequent words retained (majority baseline = 51.9%).

	OQP+GOV	OQP	GOV
With 500 most frequent features retained			
English	78.5	81.7	72.6
French	76.6	78.3	71.2
With 500 most frequent features removed			
English	76.3	73.5	71.9
French	75.0	76.1	61.9

4.2.3 *Including the other opposition parties*

Another way to see whether the classifier is more sensitive to party status than to ideology is to muddy the ideological waters by including the left-wing parties, which were in opposition in both Parliaments, in the analysis. If the classification were truly ideological, lumping these parties in with the other conservative (36th Parliament) or liberal (39th Parliament) opposition parties would markedly degrade the performance of the classifier. On the other hand, if party status is what matters, there should be little effect in doing so as the opposition parties will be more or less indistinguishable. We carried out this experiment on the English data with frequent words retained.

The results are shown in Table 8. They should be compared with the liberal/conservative results for the same Parliament and same processing method, shown in the first lines of Table 1 (96.9%, 83.3%) and Table 3 (88.3%, 72.3%). There is almost no degradation of performance on the 36th Parliament; for the 39th Parliament, there is a noticeable drop (10.12 percentage points) for the question period, but little for the debates.

Table 8. Accuracy (%) of classification of government and opposition (all parties) on English text of the 36th and 39th Parliaments with the 500 most frequent words retained (majority baselines = 51.5% and 59.4% respectively).

	OQP	GOV
36th	95.6	82.6
39th	78.2	70.9

4.3 Discussion

The results seen in Sections 4.2.1–3 are consistent with the hypothesis that the SVM bag-of-words classifier is sensitive not to expressions of ideology for which party membership is a reasonable proxy, but rather to expressions of attack and defence, opposition and government. When we train on one parliament and test on another in which party roles have been interchanged, the performance of the classifier completely disintegrates; the degradation is far worse than can be explained merely by the difference between the two parliaments in the vocabulary of the current topics of discussion. Some features that are indicative of each party 'swap sides' with the change of government. And combining ideologically inconsistent opposition parties in the classifier does not in most cases seriously degrade its performance.

5. Classification based on the emotional content of speeches

Recall that our feature analysis of the 36th Parliament showed that liberal members tended to use words that convey a more positive sentiment than those used by conservatives. This suggests that it might be possible to distinguish parties or ideologies (solely) by the emotional content of their speeches. Indeed, researchers such as James Pennebaker have made something of an industry of interpreting politicians from a statistical analysis of their use of a single category of words. For example, during the 2008 U.S. presidential election, Pennebaker (2008) wrote:

> Over the last few years, some have argued that the use of negations (e.g., *no, not, never*) indicate [*sic*] a sign of inhibition or constraint. Low use of negations may be linked to impulsiveness. … Across the election cycle, Obama has consistently been the highest user of negations – suggesting a restrained approach – where as [*sic*] McCain has been the lowest – a more impulsive way of dealing with the world.

Similarly, Pennebaker concluded that McCain's greater use than Obama of the first-person singular (*I, me, my*) signalled a likely greater openness and honesty.[10]

In the context of our results above, the questions we ask are not just whether liberals can be distinguished from conservatives in the Canadian Parliament

10. The validity of this kind of analysis is discussed and defended by Pennebaker et al. (2007a). But Pennebaker (2008) also writes: "No one should take any text analysis expert's opinions too seriously. The art of computer-based language analysis is in its infancy. We are better than tea-leaf readers but probably not much."

merely by the emotional content of their speeches, but also, if so, whether the feature actually discriminates ideology (in line with the stereotype of happy liberals, dour conservatives) or is again confounded by the parties' status in the Parliament.

5.1 Method and data

To test these questions, we used Pennebaker et al.'s (2007b) software *Linguistic Inquiry and Word Count (LIWC2007)*. LIWC counts the proportion in a text of particular words and word stems in over 60 categories, including linguistic properties (pronouns, adverbs, prepositions, etc), psychological denotation (positive emotion, negative emotion, etc), and various topics (work, money, religion, etc); it does not, itself, provide any interpretation of the counts.

For these experiments, we used the English speeches of the oral question periods and debates of the 36th and 39th Parliaments, excluding MPs who spoke very little. This gave us a dataset of documents for 200 MPs (121 liberals, 79 conservatives) in the 36th Parliament and 220 MPs (125 conservatives and 95 liberals) in the 39th Parliament. First, we ran LIWC on this data, which gave us a 64-component vector for each document, each component being the value that LIWC computed for the document for one of its categories. We then performed classification experiments on the data (with five-fold cross-validation) using this 64-component representation of the documents, in order to see whether positive and negative emotion were among the top discriminating features for liberals and conservatives, respectively. Then we repeated the classification, using *only* positive emotion and negative emotion (referred to as POSEMO and NEGEMO) as features. Finally, we performed a third experiment, in which affect was reduced to a single feature, the amount by which the positive emotion in the text exceeded the negative (i.e., POSEMO minus NEGEMO); this representation does not distinguish a completely unemotional text from one that contains emotion of each polarity in equal amounts.

5.2 Results

Table 9 shows the results of these experiments. In the first experiment, with 64 features, the accuracy for both datasets was equal to the majority baseline, because all MPs were classified as members of the majority party! In contrast, using only POSEMO and NEGEMO, either as two features or as a single feature, yielded a substantial improvement of up to 20.5 percentage points over the baseline (a relative

Table 9. Accuracy (%) of classification by party using LIWC features for English text of the 36th and 39th Parliaments' oral question period (OQP) and debates (GOV) (majority baseline = 60.5% and 56.8% respectively).

	36th		39th	
	OQP	**GOV**	**OQP**	**GOV**
64 features	60.5	60.5	56.8	56.8
POSEMO and NEGEMO	80.5	79.5	73.1	55.0
POSEMO minus NEGEMO	81.0	78.5	72.2	59.1

error reduction of 51.9%) for the 36th Parliament and 16.3 points for the oral question periods of the 39th. However, performance remained around baseline for the debates of the 39th Parliament.

Nonetheless, a feature analysis confirmed that in the 36th Parliament, positive emotion was among the top five liberal features and negative emotion was among the top ten conservative features, whereas in the 39th Parliament, positive emotion was the fourth feature for conservatives in oral question periods and sixth in debates, whereas negative emotion was eighth and tenth respectively for liberals. Hence, we can see that positive emotion is a characteristic of members of the governing party, and negative emotion is a characteristic of members of an opposition party; again, party status confounds ideological classification. The result of the classifier on all 64 features may be explained by the fact that no LIWC category had a *significant* impact on the classification. In other words, even though some LIWC categories were discriminating features for liberals and others were discriminating features for conservatives, the overall difference between the two groups was so slight that without feature selection the resulting classifier simply labelled all test instances as belonging to the majority class. This seems to be the case also for POSEMO and NEGEMO by themselves in debates in the 39th Parliament.

6. Third set of experiments: European Parliamentary data

If our 'ideological classifier' is in reality sensitive to government and opposition, then this effect should disappear when it is applied to data in which there is no government or opposition per se, but merely position-based debate with a more or less equal amount of attack and defence on both sides. Such a situation may be found in the European Parliament, in which a left–right ideological division dominates government–opposition divisions (Hix et al. 2007).

Our goal here is thus very similar to one of the tasks of the 2009 DEFT text mining challenge[11] (DEFT 09): classification by political group of speeches by Members of the European Parliament (MEPs). The DEFT corpus consisted of speeches from 1999 to 2004 by MEPs belonging to the five largest groups. Three teams attempted this task, but two declined to share their results. The remaining team, from the University of Montreal (Forest et al. 2009), reported an *F*-measure of about 0.33 on multiclass classification, which the organizers described as "mediocre" as the random baseline accuracy for the corpus was about 28% (Grouin et al. 2009: 49). We attempted both binary classification of left-wing versus right-wing MEPs, and multiclass classification of MEPs from the five largest groups, as in the DEFT task.

6.1 Data

We used English data from the proceedings of the European Parliament as our corpus.[12] Ranging from 2000 to early 2010, it was almost a strict superset of that used in the DEFT task. However, the data used in the DEFT task had been stripped of any explicit references to groups. Thus, tokens such as *PPE, Christian-Democrat,* and *United Left,* were all replaced with an anonymous tag. We understand that this was because, in the DEFT task, the organizers had human judges attempt a classification on the same data for comparison, and phrases such as *As vice-chairman of the PPE-DE group, I…* were presumably considered too much of a giveaway to a human reader. By contrast, we left all group names in place in our data. In Section 6.4.3 below, we will discuss the effect that anonymization has on classification.

6.2 Method

The choice of how to organize the raw text into vectors proved to be a key one. Our first approach was, for each MEP, to concatenate all of their utterances and consider that to be one document, as we did for the Canadian Parliament. This

11. DEFT (DÉfi Fouille de Textes) is an annual challenge and evaluation conference for researchers in text mining and classification. Each year, one or more tasks related to text mining are set, and training and test corpora are provided; research teams compete to get the best results. Results and methods are then discussed at the conference.

12. The data was collected and marked-up in XML by Dr Maarten Marx of the University of Amsterdam, who kindly made it available to us; see Marx and Schuth (2010) for details of the XML markup.

contrasted with the approach taken in the DEFT challenge, in which each individual speech remained a separate document. With our concatenation policy, however, we achieved accuracy only slightly above a random-guessing baseline. However, we observed that the amount of text we had per MEP varied widely, from a few hundred words to tens of thousands of words. Yet each MEP's document was being turned into a vector that affected the classifier equally, contradicting the natural notion that a document should have an affect on the classifier commensurate with its size. We rectified this by dividing each MEP's concatenated utterances into a number of equal-sized documents. We experimented with different document sizes, and found that it had a marked effect on accuracy (as shown in our results below).[13] The sizes that we used begin with 267 words, which was the average document length in the DEFT challenge.

As features, we used log *tf-idf*–weighted word types with words appearing in fewer than five documents removed (though we experimented with a variety of pre-processing methods, none of which had a profound effect on our results). We used SVMs for binary classification and SVM-multiclass for multiclass classification. All of the results presented below are the averaged results of five-fold cross-validation.

Binary classification: In performing binary classification, we were first faced with the task of meaningfully splitting the groups involved into left-wing and right-wing, a task that was further complicated by changes in groups and their names over the ten-year study period and by inconsistencies in identification of the groups in the data (e.g., *Greens, Verts*).[14] From descriptions of the groups, we classified as either broadly left or right 15 of the 18 affiliations observed in the data,[15] which we then grouped into the ten bins shown in Table 10.

Multiclass classification: We followed the example of the DEFT task in using only the five largest groups for multiclass classification (see Table 10), excluding the smaller right-wing groups. In multiclass classification, we found that tuning the error cost *C* on a logarithmic range of values was especially important, and that our best results were achieved with *C* on the order of 10^9.

---

13. If an MEP spoke significantly less than the document size, they were discarded from the data. Even with the highest value for document size (6666 words), this depleted the data by only 2%. In our earlier Canadian experiments described above, we discarded small documents but we did not subdivide large ones.

14. Group abbreviations usually appeared in French – e.g., *PSE* rather than *PES* for the Party of European Socialists – even in the English data. Here, we use the predominant label.

15. Omitted were the *non-inscrits* (independents), the Technical Group of Independents (a group described as politically heterogeneous), and the Alliance of Democrats and Liberals (ALDE) (described variously as conservative, liberal, or as centrist).

Table 10. European political groups as clustered, ordered from left-wing to right-wing. For the purposes of binary classification, groups above the centrist group ALDE are considered left-wing (L), and all groups below are considered right-wing (R). Asterisks mark the five largest groups, which were used in the multiclass classification experiments.

Group	Speakers in corpus	Description	L/R
*NGL	104	Communist / far-left	L
*PSE	446	Social democrats	L
*Greens	114	Green	L
*ALDE	195	Liberal / centrist	–
*PPE	571	Conservative / Christian democrat	R
ECR	41	Conservative	R
EDD	75	Eurosceptic	R
UEN	75	National conservatism	R
EFD	22	Eurosceptic, national conservatism	R
ITS	18	Far-right nationalist	R

6.3 Results

Table 11 shows the accuracy of binary left–right classification with varying document sizes. The ten words most characteristic of each class are shown in Table 12.

Table 13 shows the accuracy of multiclass classification for varying document sizes. The confusion matrix for multiclass classification is shown in Table 14; it reflects a limited subset of the data, chosen so that each group was equally represented.[16] Table 15 shows the ten words best characterizing each of the five classes.

Table 11. Precision, recall, and accuracy (%) of left–right classification on speech in the European Parliament, with varying document sizes. Baseline accuracy (more frequent class) is 50–51%, varying slightly with document size.

Document size (words)	Precision	Recall	Accuracy
267	62.6	65.2	62.3
833	67.6	70.1	67.4
1667	69.9	71.9	69.8
3333	72.9	77.5	73.9
6666	77.6	81.3	78.5

16. Note that the confusion matrix reflects a sample of 1350 documents from the almost 3000 that we considered. We chose this sample so that each group had an approximately equal number of documents. (Classification of the full set of documents tended to favour the groups which were heavily represented, thus obscuring the measure of ideological similarity we were looking for.)

Table 12. The top 10 English words characterizing each class in left–right classification of speech in the European Parliament.

Rank	Left-wing	Right-wing
1	socialist(s)	subsidiarity
2	unions	christian
3	pse	strasbourg
4	employees	competitiveness
5	greens	healthy
6	scotland	prosperity
7	gender	democrats
8	equality	competitive
9	supports	communist
10	myself	truth

Table 13. Accuracy (%) of five-way multiclass classification of speech in the European Parliament by political group, with varying document sizes. Baseline accuracy (most frequent group) is 38–39%, varying slightly with document size.

Document size (words)	Accuracy
267	44.0
833	48.0
1667	52.7
3333	56.2
6666	61.8

Table 14. Confusion matrix for multiclass classification of speech in the European Parliament by political group. Column headings are our classifications, rows are true affiliations. Boldface indicates correct classifications; italics indicates incorrect classification of a group as an ideologically adjacent group. Shaded cells show confusion between the PPE and the PSE.

	NGL	PSE	Greens	ALDE	PPE	Total
NGL	**204**	*17*	36	9	10	276
PSE	*16*	**136**	*20*	34	71	277
Greens	20	*25*	**153**	*30*	16	244
ALDE	3	39	*14*	**170**	*50*	276
PPE	3	65	9	*41*	**159**	277
Total	246	282	232	284	306	1350
Accuracy (%)	73.9	49.0	62.7	61.5	57.4	61.8

Table 15. The top 10 English words characterizing each group in multiclass classification of speech in the European Parliament.

Rank	NGL	PSE	Greens	ALDE	PPE
1	confederal	socialist	greens	liberal	christian
2	nordic	socialists	alliance	liberals	subsidiarity
3	military	pse	nuclear	eldr	conservatives
4	unemployment	institution	ngos	democrat	morning
5	profits	eplp	basque	alde	wrong
6	occupation	balanced	scotland	alliance	competitiveness
7	nato	interinstitutional	comments	rapporteurs	healthy
8	liberalization	millennium	planes	obvious	competitive
9	yugoslavia	repeatedly	ale	china	communist
10	militarisation	portuguese	conflict	7	phenomenon

6.4 Discussion

6.4.1 *Comparison with DEFT results*

With equal average document length, our multiclass accuracy was only marginally better than the DEFT results of Forest et al. (2009) (about 5 percentage points over baseline, rather than 2 points). However, as document size was raised to a maximum of 6666 words, accuracy increased steadily, up to 61.8%, about 23 points over baseline. This suggests that the average DEFT size of 267 words is simply an insufficient size for bag-of-words-based methods over such a noisy corpus.

6.4.2 *Relative difficulty of classification tasks*

The accuracy of multiclass compared to binary classification suggests that associating a speech with a specific group is not much harder than just classifying it as left or right. This may, more than anything else, speak to the composition of the European Parliament. Hix, Noury, and Roland (2007) suggest that, rather than falling at some point on a line from left to right, European Parliament groups can be placed in a space where the primary dimension "is the traditional left–right axis and the second dimension is a mixture of attitudes towards European integration (in favour and against) and government–opposition status in the EU" (p. 217). In addition, 'green' sentiment, while often lumped in with liberalism, is not quite a strict subset thereof (and implies a completely different vocabulary). These multiple dimensions complicate the task of binary ideological classification.

The confusion matrix for multiclass results (Table 14) may shed some light on the relative ideological distances between groups. As we would wish, confusion is for the most part clustered around ideologically similar groups. Because

groups are arranged from left to right in order of ideology, this fact is reflected by the tendency of confusion to cluster around the diagonal. The most surprising result is the high amount of confusion between PSE (a socialist group) and PPE (Christian democrats, the most conservative group we considered) (shaded cells). This may be because these two groups had perhaps the least coherent feature lists (see Section 6.4.3 below). It may be significant that the two most accurately classified groups, NGL and the Greens, also had the most subjectively coherent feature lists.

6.4.3 *Discriminative features*

A few trends emerge from the lists of the top features for each group. The most obvious is that MEPs tend to talk about their own groups. Hence, the top feature for the Greens is *greens*, the top two features for the PSE are *socialists* and *socialist*, and so on. This contrasts with Canadian MPs who we found (Section 3.5 above) tend to talk about their opponents more than themselves. This striking difference demonstrates the domain-specificity of the features learned by the classifier. We do however find some instances of MEPs talking about their opponents, most notably the appearance of *communist* in the top 10 features of the PPE, the most right-wing group we looked at. As might be expected, the contexts in which PPE MEPs actually used the word were highly negative, phrases such as *communist tyranny*. But clearly, whether MEPs talk about themselves or their opponents, the names that each group tends to utter are important discriminators. Thus we see a second reason why the results of Forest et al. (2009) were so poor; anonymization of the groups removes crucial discriminators.

Some of the top feature lists are highly coherent with respect to the issues of concern to the group. For example, among the top 50 features for the Greens we find *nuclear*, *organic*, *contaminated*, *ecological*, *toxic*, *culling*, and *depleted*. The top 50 features for NGL, the most left-wing of the groups, included *wages*, *unemployment*, *capitalist*, *wage*, *inequality*, and *poverty*. The top features for right-wing PPE are less coherent, though as Diermeier et al. (2007) found of right-wing U.S. senators, there tended to be a focus on cultural and moral issues: *christian*, *moral*, *conscience*, *faith*, and *euthanasia* all appear in their top 100 features.

Some trends that the classifier seems to pick up on aren't overtly ideological, and indeed hint at the language of attack and defence. In the case of centre-left group PSE, the classifier seems tuned to the language of felicitation, with words like *wholehearted*, *congratulations*, *congratulating*, *impressed*, *proud* and *achievement* all in their top 50, whereas the centrist group ALDE seems to be associated with censorious language: *accountability*, *needless*, *shameful*, *shame*, *breaches*.

7. Conclusion

Our results cast doubt on the generality of the results of research that uses words as features in classifying the ideology of speech in legislative settings – and possibly in political speech more generally. Rather, the language of attack and defence, of government and opposition, seems to dominate and confound any sensitivity to ideology. Such research therefore reduces in effect to the classification of support or opposition, much as in the linguistic component of the work of Thomas et al. (2006) described in Section 1 above. However, even if our classifiers are construed as distinguishing support from opposition, our results are much more accurate than those of Thomas et al., even though we did not use any explicit component for detecting agreement or disagreement between individual speakers. This may be partly attributed to one of the differences between Canadian and U.S. politics: Canadian parties have strong party discipline, and agreement between speakers may be reliably inferred from shared party membership.

Our results contrast with the conclusions of Diermeier et al. (2007), who argue from their own results that speakers' words in debates in the U.S. Congress are "expressions or representations of an underlying belief system". Again, political differences might be a partial explanation of the difference. Perhaps the weak party discipline of the U.S. and the separation of the Congress from the Executive branch motivates greater attention to ideological substance in debates than does the Canadian (Westminster-style) system in which an explicit governing party, including the head of government and all cabinet ministers, is represented as such in the legislature. This possibility is supported by our results from European Parliamentary data. But this is speculation; our results have demonstrated a confound that must be taken into account in research on ideological classification of speech in any context.

References

DEFT (2009). Actes de l'atelier de clôture de la cinquième édition du DÉfi Fouille de Textes, Paris, 22 June 2009.

Diermeier, D., J.-F. Godbout, B. Yu and S. Kaufmann. 2007. Language and ideology in Congress. *Annual Meeting of the Midwest Political Science Association*.

Forest, D., A. van Hoeydonck, D. Létourneau and M. Bélanger. 2009. Impacts de la variation du nombre de traits discriminants sur la catégorisation des documents. *Actes de l'atelier de clôture de la cinquième* édition *du DÉfi Fouille de Textes,* Paris, 22 June 2009, pp. 75–88.

Germann, U. 2001. Aligned Hansards of the 36th Parliament of Canada. Available at www.isi.edu/natural-language/download/hansard/

Greene, S. 2007. Spin: Lexical semantics, transitivity, and the identification of implicit sentiment. PhD thesis, University of Maryland, College Park.

Grouin, C., B. Arnulphy, J.-B. Berthelin, S. El Ayari, A. Garcia-Fernandez, A. Grappy, M. Hurault-Plantet, P. Paroubek, I. Robba and P. Zweigenbaum. 2009. Présentation de l'édition du D'Efi Fouille de Textes (DEFT'09). *Actes de l'atelier de clôture de la cinquième* édition *du DÉfi Fouille de Textes,* Paris, 22 June 2009, pp. 35–50.

Hix, S., A. G. Noury and G. Roland. 2007. *Democratic Politics in the European Parliament*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511491955

Jiang, M. and S. Argamon. 2008. Political leaning categorization by exploring subjectivities in political blogs. *Proceedings, 4th International Conference on Data Mining (DMIN 2008)*, pp. 647–653.

Lin, W., T. Wilson, J. Wiebe and A. Hauptmann. 2006. Which side are you on? Identifying perspectives at the document and sentence levels. *Proceedings of the 10th Conference on Natural Language Learning (CoNLL-X)*, pp. 109–116.

Marx, M. and A. Schuth. 2010. DutchParl: The parliamentary documents in Dutch. *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010)*, Malta.

Mullen, T. and R. Malouf. 2006. A preliminary investigation into sentiment analysis of informal political discourse. *Proceedings of the AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 159–162.

Pennebaker, J. W. 2008. The meaning of words: Obama versus McCain. Weblog, 12 October 2008. Available at wordwatchers.wordpress.com/2008/10/12/

Pennebaker, J. W., C. K. Chung, M. Ireland, A. Gonzales and R. J. Booth. 2007a. *The Development and Psychometric Properties of LIWC2007*. Available at www.liwc.net/LIWC2007 Language-Manual.pdf

Pennebaker, J. W., C. K. Chung, M. Ireland, A. Gonzales and R. J. Booth. 2007b. Linguistic inquiry and word count (LIWC2007). Available at www.liwc.net

Riabinin, Y. 2009. Computational identification of ideology in text: A study of Canadian parliamentary debates. MSc paper, Department of Computer Science, University of Toronto, January 2009. Available at www.cs.toronto.edu/compling/publications.html

Thomas, M., B. Pang and L. Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pp. 327–335. DOI: 10.3115/1610075.1610122

Yu, B., S. Kaufmann and D. Diermeier. 2008. Classifying party affiliation from political speech. *Journal of Information Technology in Politics* 5(1), pp. 33–48. DOI: 10.1080/19331680802149608

CHAPTER 6

Sentiment analysis in parliamentary proceedings

Steven Grijzenhout, Maarten Marx, and Valentin Jijkoun
University College London (UCL) / University of Amsterdam

This chapter addresses the question whether opinion-mining techniques can successfully be used to automatically retrieve political viewpoints from parliamentary proceedings. Two specific preprocessing tasks were identified and systematically evaluated: automatically determining subjectivity in the publications and automatically determining the semantic orientation of the subjective parts. A corpus of recent parliamentary proceedings was collected and a gold standard annotation was created on both subjectivity and orientation. Following this, a number of models based on subjectivity lexicons and machine-learning algorithms were evaluated. Machine-learning algorithms perform best, but methods based on subjectivity lexicons also provide promising results. Based on these results we can conclude that opinion-mining techniques applied to political data score just as well as the state of the art in other more traditional domains of opinion mining like product reviews and blogs.

1. Introduction

Opinion mining is a recent discipline concerned with automatically determining the opinion a text expresses (Pang and Lee 2008). Opinion mining and sentiment analysis are terms that are used more or less synonymously in the literature, and we will also use them interchangeably. In this paper, we evaluate whether opinion-mining techniques can successfully be applied in political text analysis. As data we use the verbatim transcriptions of the plenary meetings in the Dutch House of Representatives. These documents are an important source of information on the position of political parties and individuals in the political arena. Our research concerns the following central research question: Can opinion-mining techniques be used to automatically retrieve political viewpoints from parliamentary proceedings? To answer this question (albeit indirectly), we evaluate whether opinion-mining techniques are appropriate methods to analyze political data. This approach is warranted as opinion-mining techniques have been tested elaborately

on product reviews and blogs, but not yet on political data. We will show that the data format of parliamentary proceedings is well suited to do sentiment analysis. We then proceed to evaluate the two key steps in sentiment analysis: determining subjective passages, and in these passages, determine the semantic orientation, that is: do they express positive or negative attitude?

The paper is organized as follows: Section 2 provides background information on the various approaches to sentiment analysis; Section 3 describes the data; Section 4 describes the task and the experiments we have done, and presents the results; Section 5 reviews methods and results and compares them with other research in this field.

2. Background

This research is part of a set of four research areas: opinion search, opinion mining, topic mining and recent research regarding Dutch parliamentary proceedings. In this section these research areas are discussed briefly.

1. *Opinion search*. Opinion search is a relatively new branch of research. It aims to enable users to search for opinions on any object (Liu 2007). Here, the entity "object" is used to point to different concepts including products, persons, happenings or topics. Therefore opinion search can be helpful for a broad range of applications, including review-related websites, blogs, business intelligence, government intelligence and politics. Most research covers opinion search applications in the context of blogs (web-logs) and review-related websites.

2. *Opinion mining*. Opinion mining concerns analyzing the opinion a text expresses. Motivated by real-world applications researchers have considered a wide range of problems in this area (Pang and Lee 2008). Esuli and Sebastiani (2006) have organized these problems into three categories:

1. Determining *subjectivity*: the problem of determining whether a given text passage has a factual nature or expresses an opinion.
2. Determining *orientation* (also called *polarity*): the problem of determining whether a given subjective text expresses a positive or negative opinion.
3. Determining the *strength of orientation*: for example, weakly positive or strongly negative.

A closely related task is extracting information on *why* the topic or product in the text is considered positive or negative (Pang and Lee 2008). Other research problems include automatically determining the political color of a text, for example, liberal or conservative (Mullen and Malouf 2006). The three categories identified

by Esuli and Sebastiani (2006) account for the majority of the research in opinion mining. In this paper we evaluate algorithms for the first two categories: automatically determining subjectivity, and determining the orientation of the subjective passages.

3. *Topical sentiment analysis.* Here the goal is not only to determine that a passage expresses a certain attitude, but also to determine the object of the attitude. Note that in many cases this is known from other sources than the passage (e.g., it is listed separately in a product or movie review). A common approach is to apply a categorization algorithm to a text and then perform a sentiment analysis (Osman and Yearwood 2007).

4. *Parliamentary proceedings.* More and more large historical corpora of parliamentary proceedings become available for research. Examples include the British Hansard and the Dutch Parlando. These are two search engines which provide the digitized parliamentary proceedings in a machine-readable XML format. These corpora contain a wealth of information for historical political analysis but digital sustainability and good access to them remains a research challenge (Marx et al. 2010).

3. Data

We have tested our algorithms on the verbatim transcripts of the plenary meetings of the Dutch House of Representatives (Tweede Kamer: Plenaire vergaderingen (n.d.)). These are available in a variety of technical formats (PDF, Word, HTML, XML). All of these, except the XML format, are meant for human reading and not for machine processing. In this respect the Dutch data is typical for parliamentary proceedings (Marx and Schuth 2010).

It is a technical challenge to transform some of these formats into useful machine processible formats, especially for the older scanned and OCRed material (Marx and Schuth 2010). In order to determine political viewpoints from these texts, the following information, easily detected by humans but not by machines, is needed: for each word in the text we need to know whether it was spoken or not. If so, by whom, the role or function of the speaker, when, and in which context. For sentiment analysis, we also need a reliable segmentation of the text into words and paragraphs.

These desiderata were best met when using the HTML version of the proceedings data. These are available from the Dutch parliament directly, one day after each meeting, as a draft version (from www.tweedekamer.nl). The transcripts were downloaded and automatically transformed into the XML format described in Marx and Schuth (2010). An example is provided in the Appendix.

4. Assessing subjectivity and orientation

In this section, we present several sentiment analysis techniques, apply them to our dataset, and systematically evaluate their quality using a manually annotated corpus. We perform the first two classification tasks according to the schema of Esuli and Sebastiani (2006). First we determine whether a text is subjective or objective. Second, for the subjective texts, we determine whether they have a positive or negative orientation. Before we start with the classifiers we first must decide on the level of detail on which classification is done and on how to create the gold standard corpus.

4.1 Classification level

Before classification can commence, the level at which it will be conducted needs to be chosen. Different levels are used in the literature:

- *Document level* (Yu and Hatzivassiloglou 2003): whole documents are labeled. For example, a document can have an overall orientation that is classified as positive.
- *Block level* (Osman and Yearwood 2007): the text is cut into several blocks and each block is labeled independently. This is most often used in unstructured data like blog pages.
- *Paragraph level* (Kamps and Marx 2001): each paragraph is labeled.
- *Sentence level* (Riloff and Wiebe 2003; Wilson et al. 2003; Furuse et al. 2007: each sentence is labeled.
- *Word level* (Yu and Hatzivassiloglou 2003; Kim and Hovy 2005; McKeown and Hatzivassiloglou 1997): individual words are labeled.

Classification at document level and word level is unsuitable for identifying political viewpoints in parliamentary proceedings. Document level classification means a whole meeting is treated as an individual entity, and marking it will give no particular views of individual parties or political persons. It is too general to be of value. In contrast, classification at the word level is too detailed, and will not contain enough contextual information to connect sentiment to a particular viewpoint or topic.

Classification at the sentence level also has problems with contextual information since individual sentences will contain references to adjacent sentences and topics. For example, the sentence 'That is okay.' contains an opinion, but we do not know what the opinion is about. Arguments containing a viewpoint are often expressed in multiple sentences. This leaves us with the choice between either

block level or paragraph level classification. Since a paragraph is considered a natural block, this has been considered most appropriate. The source data was already split into paragraphs (using the p-element in HTML), so no further processing was needed.

4.2 Gold standard corpus

Evaluation of opinion retrieval algorithms mostly relies on a comparison with human annotations from the same corpus (Ku et al. 2006.; Osman and Yearwood 2007). To evaluate the performance of the algorithms on the Dutch parliamentary proceedings, a gold standard was developed. We aimed at annotating around a thousand paragraphs. For efficiency reasons, the paragraphs were extracted from as few documents as possible. We randomly chose two meetings (March 5th and April 21st, 2009) which contained enough (1201) paragraphs. Paragraphs spoken by the chairman were not annotated because the chairman does not take part in the discussions on political issues, but instead tries to keep the meetings on track.

The first task was to annotate whether a paragraph contains an opinion or not. If there is an opinion present, the paragraph is considered subjective. Otherwise the paragraph is considered objective. Examples of objective and subjective paragraphs are given in the Appendix. Two human annotators were used, both with Dutch as their mother tongue. The paragraphs were printed and split evenly between them. A face-to-face explanation of the intention of the research and their task of annotating the paragraphs was provided. The annotators marked each paragraph as subjective or objective. This was judged by reviewing each individual paragraph against a definition of subjectivity. The definition of subjectivity that was used is based on the literature (Banea et al. 1999; Kim and Hovy 2004; Riloff and Wiebe 2003; Wiebe and Riloff 2005; Wiebe et al. 1999), the *Compact Oxford English Dictionary* definition of *opinion*, and the Dutch *Van Dale online dictionary*'s definition of *mening*)(Van Dale online dictionary). Our definition is:

> If the primary intention of a piece of text is an objective presentation of material that is factual to the reporter, and does not contain a judgment or emotion, the text is objective. Otherwise the text is subjective.

The second task was to annotate the semantic orientation of each subjective paragraph. As mentioned, the orientation of a text is whether it expresses a positive or negative opinion. The same two annotators were used. This time, however, instead of splitting the paragraphs evenly between them, the two annotators individually marked all of the subjective paragraphs. Discontinuities between the annotators

were afterwards resolved via mutual consultation. To have two judgments meant that the inter-annotator agreement could be monitored (see below).

In the literature, a clear definition of positive and negative orientation is hard to find. Most of the time multiple human annotators are used to judge a corpus based on their intuition or common sense (Furuse et al. 2007; Jijkoun and Hofmann 2009). Based on research by Osgood et al. (1957) the semantic orientation on which we wish to classify the paragraphs is the evaluative factor: good/bad. They proved that this factor is the most significant influence on variation in data. A definition of orientation based on this research can be found in Turney (2001): "a phrase has a positive semantic orientation when it has good associations and a negative semantic orientation when it has bad associations". Turney and Littman (2003) also distinguish between positive evaluation (e.g., praise) and negative evaluation (e.g., criticism) respectively. From these sources, the following definition to classify this binary orientation was formulated, leaning heavily on Osgood et al. (1957). The annotators were instructed to use this definition in their task:

> A text has a positive orientation when it has good associations, or contains a positive evaluation (e.g., praise). The text has a negative orientation when it has bad associations or contains negative evaluations (e.g., criticism).

Corpus statistics. In the transcripts of the meetings of March 5th and April 21st, 2009, a total of 1201 paragraphs were annotated, of which 590 (49.1%) were annotated as subjective. Out of these 590 subjective paragraphs, 251 (42.5%) were annotated as positive, and 339 (57.5%) as negative. Because two annotators were used, inter-annotator agreement could be calculated. The overall agreement is 71.4%. There was hardly any difference in agreement between the positive and the negative paragraphs (175/251 = 69.5% and 246/339 = 72.4%, respectively). Cohen's κ is 0.423.

Conclusions. Because of the use of definitions, the annotation tasks were easy to explain to the annotators. Also, because strict definitions were used, the annotators did not need to have specific domain knowledge. According to some, an analytic definition of opinion is impossible (Kim and Hovy 2004). Still, even with the strict instructions, inter-annotator agreement was low: overall agreement on semantic orientation between the two annotators was 71.4% and $\kappa = 0.423$. These rather low values are however common for sentiment analysis tasks: e.g., (Kim and Hovy 2005) classified 174 sentences by three annotators and found a pair-wise agreement of 73% and a kappa value of 0.49.

4.3 Automatically determining subjectivity

We now describe and evaluate a number of algorithms for determining subjectivity of texts. Technically, these algorithms are binary classifiers: for each input they decide whether it is subjective or not. Algorithms for such a task fall into two categories: based on handcrafted rules, or (machine) learned from examples. Algorithms in the first category make use of so called subjectivity lexicons. We will evaluate two algorithms based on lexicons and three based on machine learning.

Algorithms based on subjectivity lexicons

In this approach, the focus is on the number of occurrences of each term. The exact ordering of the terms in a text is not important (Manning et al. 2008). Most often the individual words are given a certain subjectivity score based on a set of opinion words (the subjectivity lexicon) (Ding and Liu 2007). The models then present a way to calculate the subjectivity of the whole text based on the individual collection and frequency of these words. We discuss two models from Kim and Hovy (2005) which implement this idea.

Model 1 counts the total valence score of all words in the paragraph. The basis of this model is that paragraphs dominated by words considered to be subjective tend to be opinion bearing. Individual words in the paragraph are extracted and given a score of 0, 1 or 2, in which a score of 2 is considered to be very subjective and a score of 0 not subjective. A Dutch sentiment wordlist developed by Jijkoun and Hofmann (2009) was used to rate the words. Words not present in the wordlist are considered to be not subjective and have been given a score of 0. A cut-off threshold had to be selected in order to determine when a paragraph is judged to be subjective or objective. Experimentation has been conducted with threshold values between 0 and 20.

Model 2 checks the presence of a single strong valence word. The assumption underlying this model is that the presence of one strong valence word is enough to indicate subjectivity. The Jijkoun and Hofmann sentiment lexicon (2009) is used and a cut-off threshold is set to determine at which score a paragraph is considered to be subjective. Because the wordlist by Jijkoun and Hofmann (2009) contains scores of 0, 1 and 2, where 0 indicates neutrality, the performance of the algorithm is evaluated on cut-off thresholds of 1 and 2.

Machine-learning algorithms

Machine-learning algorithms differ from models based on subjectivity lexicons in that they automatically train themselves to classify the data. The methods we use are called supervised methods because a labeled data set is needed to train

the classifier. For this we use our gold standard. The following machine-learning algorithms are used (Manning et al. 2008):

- NaiveBayes;
- IBk nearest-neighbour, with k = 1;
- Support Vector Machine (SVM) SMO;
- ZeroR (as the baseline).

The toolkit Weka 3.6.1 is used to train and evaluate the machine-learning classifiers (WekaWiki: Primer (n.d.)).

Results

The performances of the algorithms are based on accuracy, precision, recall and F-measure (Manning et al. 2008). The machine learning algorithms are evaluated using tenfold cross-validation.

As described above, Model 1 uses a cut-off threshold above which the text is classified as subjective. The performance of the model at different cut-off thresholds can be found in Table 1. The highest scores are in bold font. Because the aim of the subjectivity classification is to retrieve paragraphs that are subjective, we are interested mostly in the results of the TRUE-class. (The TRUE class consists of the set of subjective paragraphs.) A higher threshold will lower recall, and will increase precision. This is as expected, as a higher cut-off threshold will include less subjective markers (Kim and Hovy 2005). The table should be interpreted as follows: A threshold of > n means that a paragraph is classified as subjective if the sum of the valence scores in the paragraph is larger than n. Thus the threshold "> 0" means that a paragraph is subjective if it contains at least one word with subjectivity score 1. We see that with threshold ">1", we find 93% of all subjective paragraphs (recall = .929) and that 55% of those classified as subjective are indeed subjective (precision = 0.550).

Model 2, based on Kim and Hovy (2005), also uses a cut-off threshold. Here the thresholds have a different meaning: threshold 1 means that the paragraph contains at least one subjective word (thus with score 1 or 2); threshold 2 that it contains at least one word with score 2 (a highly subjective word). The results are in Table 2. Threshold 1 has the best tradeoff between precision and recall (F = .677). Note that Model 2 with threshold 1 is exactly the same as model 1 with threshold > 0: both classify a paragraph as subjective if it contains at least one subjective word.

The best results of the two models and the results of the machine-learning algorithms are shown in Table 3. In reaching these results, experiments were conducted for smoothing them out, for example, all words in the paragraphs were converted to lowercase. These experiments did not improve results significantly. In fact, converting the paragraphs to lowercase even deteriorated results.

Table 1. Results of cut-off threshold values using model 1 based on Kim and Hovy (2005).

Threshold	Results on TRUE class		
	Precision	Recall	F-measure
Baseline (all subjective)	0.491	1.0	0.659
> 0	0.521	**0.966**	0.677
> 1	0.550	0.929	**0.691**
> 2	0.570	0.854	0.684
> 3	0.591	0.775	0.671
> 4	0.605	0.686	0.643
> 5	0.636	0.615	0.625
> 6	0.653	0.546	0.595
> 7	0.671	0.478	0.558
> 8	0.671	0.395	0.497
> 9	0.684	0.337	0.452
> 10	0.689	0.275	0.393
> 11	0.695	0.232	0.348
> 12	0.688	0.186	0.293
> 13	0.705	0.146	0.242
> 20	**0.793**	0.039	0.074

Table 2. Results of cut-off threshold values using model 2 based on Kim and Hovy (2005).

Threshold	Results on TRUE class		
	Precision	Recall	F-measure
1	0.521	0.966	0.677
2	0.596	0.666	0.628

Table 3. Results of all approaches on classifying subjectivity (using optimal threshold results at K&H models for weighted results and TRUE class results).

Model	Results on TRUE class		
	Precision	Recall	F-measure
K&H model 1 (threshold >1)	0.550	0.929	0.691
K&H model 2 (threshold 1)	0.521	0.966	0.677
NaiveBayes	0.607	0.802	0.691
Ibk	0.563	0.593	0.578
SMO	0.638	0.610	0.624

Conclusions

The performance of Model 1 on the TRUE class is amongst the highest with an F-measure of 0.691: it finds almost all subjective paragraphs and classifies just over half of them correctly. How good are these results? For that we compare our scores to those obtained by Kim and Hovy (2005). However they report F measures for the weighted results of both classes.

The F-measure of Model 1 on our data is at its peak at 0.545 (threshold >6). Our implementation of Model 1 performs better on our data than the original model performed on TREC 2003 data that only achieved a F-measure of 0.425. The implementation of Model 2 also performed better on our data than the original model on TREC 2003 data, achieving an F-measure of 0.534 (threshold 2) as opposed to 0.514. Thus we can conclude that the results of the methods based on subjectivity lexicons are very promising, as they perform relatively well on political data.

From the machine-learning algorithms, NaiveBayes performs best overall with a weighted F-measure of 0.640 and an F-measure on the TRUE class of 0.691. If we select ZeroR, which predicts a class based on the mode (thus in our case: it classifies all paragraphs as objective), as baseline, NaiveBayes performs significantly better than ZeroR (*one tailed test, confidence level 0.99*). The SVM algorithm SMO also produces decent results significantly better than ZeroR (*one tailed test, confidence level 0.99*). The weighted F-measure of 0.638 comes in range of the NaiveBayes' weighted F-measure of 0.640. On classifying the TRUE class, however, NaiveBayes would still be the preferred algorithm of choice with an F-measure of 0.691 as opposed to SMO's F-measure of 0.624.

4.4 Automatically determining semantic orientation

We now evaluate algorithms which automatically determine semantic orientation. Like the algorithms determining subjectivity, these algorithms are classifiers. In line with the literature, we built binary classifiers which classify subjective paragraphs as positive or negative. Again there are lexicon based approaches and machine learning algorithms.

Algorithms based on subjectivity lexicons

The algorithm is based on the model by Edens et al. (2006) and uses the wordlist by Jijkoun and Hofmann (2009). The algortithm classifies all words as positive, negative or neutral. The scores +2 and +1 are considered positive, and −1 or −2 are considered negative. The algorithm also takes into account that two adjacent polar words of the same orientation influence each other. A factor is calculated

based on the distance between the two polar words, with a maximum distance of 10. The score of the original polar word is then multiplied by this factor. The following equation is used to calculate the new wordscore:

$$wordscore = wordscore \times (1 + \frac{10/distance}{10})$$

After all the wordscores in the paragraph have been calculated, they are added up. If the final score is above 0, the paragraph is classified as positive, otherwise the paragraph is negative.

The model based on Chesley et al. (2006) combines a subjectivity lexicon and machine learning. The expectation of this model is that the distribution of positive and negative adjectives, and positive and negative verb classes, shows regularities. Furthermore, it assumes that the orientation of adjectives can be described by the majority orientation class of their synonyms. We have implemented this model as follows:

First, a part of speech (POS) tagger (TreeTagger 3.2) is used to identify all verbs and adjectives. Next, for all adjectives, the synonyms are scraped from the website www.synonyms.net. In the original implementation by Chesley et al. (2006), Wikipedia's dictionary is used because of its coarse-grained content. We used www.synonyms.net instead because the Dutch version of Wiktionary is not sufficiently developed yet. All collected synonyms are matched against a wordlist of positive and negative adjectives. The wordlist is created by merging the adjectives of Jijkoun and Hofmann (2009) and the negative and positive adjectives collected by Kamps and Marx (2001). The majority class of the synonyms has been assigned to the adjective. The verbs are assigned to a positive or negative class based on the lexicon by Jijkoun and Hofmann (2009). As output, the model provides a list of information on each paragraph consisting of:

1. Number of positive adjectives
2. Number of negative adjectives
3. Number of positive verbs
4. Number of negative verbs

To this list, the gold standard classification on orientation belonging to the paragraph is added. Finally, using the information gathered on the paragraphs, Chesley et al. (2006) used a SVM algorithm to classify the paragraphs. They opt for the use of a SVM algorithm because they believe it to be robust for sentiment classification and handling noisy data (Mishne 2005). We use Weka's SVM algorithm SMO, but experiment with NaiveBayes, IB1, and ZeroR as well (see Table 4). NaiveBayes gave the best results.

Table 4. Results of classification on orientation using the output of the model based on Chesley et al. (2006).

Classifier	Precision	Recall	F-measure
SMO	0.567	0.581	0.560
NaiveBayes	0.597	0.602	0.599
IB1	0.553	0.558	0.554
ZeroR	0.330	0.575	0.419

Machine-learning algorithms

Again, four machine-learning algorithms are selected to represent this category.

- NaiveBayes
- IB1 nearest-neighbour
- Support Vector Machine (SVM) SMO
- ZeroR

Again, the Weka toolkit is used to evaluate these algorithms.

Results

The machine-learning algorithms are evaluated using ten fold cross-validation. Similarly to the algorithms determining subjectivity, all algorithms are evaluated on precision, recall and F-measure. Because the four features used by the model based on Chesley et al. (2006) are numeric, discretization of the data could be conducted to improve results. Experiments have been conducted with different bin sizes for each classifier. The optimal results can be found in Table 4. The following parameters were used:

- SMO: Discretization with Weka's option findNumbins. This option lets Weka choose an appropriate amount of bins.
- IB1: no discretization.
- NaiveBayes: no discretization.

NaiveBayes produces the best results on all measures: almost 60% of all classifications is correct at a recall value of .602. It is difficult to compare these results to the results found by Chesley et al. (2006) since they classify on document level of a blog post instead of on paragraph level. The results nevertheless, allow us to conclude that the SVM classifier is performing well on the data. The results support the claim of Chesley et al. (2006) and Mishne (2005) that SVM is a robust classifier for sentiment classification.

A comparison of all algorithms used to classify semantic orientation can be found in Table 5. Some smoothing experiments were conducted, but again did not improve results. As can be seen, the SVM machine learning

classifier scores best on all fronts regarding orientation classification. It is followed by NaiveBayes. Both perform significantly better than the other models used (*one tailed test, confidence level 0.99*). The combination of collecting paragraph statistics and using a machine-learning algorithm gives promising results with an F-measure of 0.599. If more characteristics are collected, performance may increase.

Table 5. Results of classifications by semantic orientation.

Model	Precision	Recall	F-measure
model based on Edens et al. (2006)	0.369	0.517	0.419
model based on Chesley et al. (2006) with NaiveBayes	0.597	0.602	0.599
Ibk	0.601	0.561	0.556
NaiveBayes	0.652	0.651	0.652
SMO (SVM)	0.677	0.676	0.677
ZeroR	0.330	0.575	0.419

5. Conclusions

The aim of this chapter was to evaluate the appropriateness of sentiment analysis techniques which are developed for blogs and product reviews, for the analysis of political texts. We first summarize our technical results and then return to their impact on the main research question.

We studied six algorithms to automatically detect opinion-bearing paragraphs. Machine-learning and lexicon-based algorithms scored about equally well. NaiveBayes performed best with a weighted F-measure of 0.640 and an F-measure on the TRUE class of 0.691. Model 1, based on subjectivity lexicons, achieved exactly the same F-measure. Next, six algorithms were studied which automatically detect the orientation of the subjective paragraphs. The algorithms classified paragraphs as either positive or negative. Machine-learning algorithms again dominated the results. NaiveBayes reached an F-measure of 0.652, but the SVM implementation in Weka called SMO performed best with an F-measure of 0.677. Both performed significantly better than the other algorithms. These results support the claim that SVM provide a solid method for sentiment classification (Chesley et al. 2006; Mishne 2005). It can furthermore be concluded that a model collecting paragraph characteristics and then classifying the paragraphs using machine-learning algorithms provides promising results.

Considering the performances of the classification algorithms, we conclude that results are approximately in line with results found in the literature. An F-measure approaching 0.7 is a common achievement, and can therefore be considered to be a respectable result. In other words, opinion-mining techniques are

suitable to automatically retrieve subjective paragraphs from texts and to annotate their orientation. This shows that today's opinion-mining techniques can be successfully applied to Dutch, political, semi-structured transcripts.

With both techniques scoring about equally well, we advise on using either lexicon based or machine-learning for political texts based on other grounds. It seems that a machine learning approach might be preferable. First, using Crowdsourcing platforms like Amazon Mechanical Turk, it becomes easy, fast, and inexpensive to create a large number of training examples. Even though the task is difficult and we cannot expect high inter-annotator agreement, this option can work fine. In usual experiments; each task is performed by five annotators. Only examples with a strong majority agreement (e.g., 4 out of 5 agree) could be used. Second, politicians are creative language users, and political language as a result changes fast. It is therefore advisable to train a learning algorithm on examples, rather than trying to capture the political language in a lexicon, especially since orientation in political data is topic-dependent. One note of caution is called for though. In case of a lack of resources, as in this study, the machine-learning algorithms are trained and tested on the same set. This might have the effect of overtraining the algorithms.

How can these results be used to retrieve political viewpoints or party positions? For retrieving a viewpoint we also need to know about which topic an opinion was given. This can be done by combining a topical paragraph classifier with the here developed sentiment classifier. Then there are still several ways to distill a party position on a topic based on a collection of paragraphs about that topic spoken by many members of that party. Thus a separate evaluation seems needed to answer the question.

The chapter by Hirst et al. (this volume) finds that lexical methods (as we use here) do not retrieve party positions but rather government versus opposition positions. Whether this also holds for opinionated paragraphs is a matter for further research. Here we can only give a first impression. We have analyzed our two days of annotated debates with respect to this question and obtained mixed results[1] (see Table 6). The debates of April 21, 2009 show an almost perfect

1. The editors of this volume made the following notes on these two days. On Tuesday April 21, the results show a perfect division between government and parliament. On Tuesdays the Dutch parliament has Question Time. This might explain the division between government and the government-parties on the one hand and parliament, more specifically the opposition parties, on the other hand. The opposition parties attack the government and the government praises itself and is praised by the government parties. March 5th, on the other hand, is a Thursday with more elaborate and detailed debates in which parties (apparently) take less clear positions. This division between question time and debates is also noted by Hirst et al. (this volume). They also find a stronger division between government and parliament during Oral Question Period than for Debates.

dichotomy between speakers from the government and parties in the government (ChristenUnie, CDA and PvdA), and the other parties. The only exception is minister Verdonk, with just 6 paragraphs. All opposition parties have more negative than positive paragraphs, while it is exactly the other way around in the other group. The other day does not have this clear division between government and opposition. On this rather gloomy day (on which even the government has almost twice as many negative praragaphs than positive), only the Christian parties (CDA, ChristenUnie and SGP) and the Greens (GroenLinks) are more positive in general.

Table 6. Results of classifications by semantic orientation.

Thursday March 5th, 2009				**Tuesday April 21st, 2009**			
Party	**POS/NEG**	**POS**	**NEG**	**Party**	**POS/NEG**	**POS**	**NEG**
ChristenUnie	1.3	4	3	ChristenUnie	3.7	11	3
CDA	1.1	27	24	Verdonk	3.0	3	1
GroenLinks	1.0	14	14	CDA	1.8	25	14
SGP	1.0	6	6	PvdA	1.6	11	7
PvdA	0.7	15	21	Government	1.2	20	17
D66	0.7	10	15	SGP	1.0	5	5
Government	0.6	33	59	PVV	0.9	12	13
VVD	0.5	16	34	GroenLinks	0.8	7	9
PVV	0.4	4	11	VVD	0.8	10	13
SP	0.2	10	41	SP	0.3	8	28
Verdonk	0.0	0	0	D66	0.0	0	0
PvdD	0.0	0	0	PvdD	0.0	0	1

EXPLANATION: We count the number of positive and negative paragraphs spoken by members of that party for each day, for each party. The second attribute is the ratio between the numbers of positive and negative paragraphs. If the ratio is higher than 1, the party is overall positive that day, below 1 indicates a more gloomy day.

Another factor that heavily influences the findings is the style of a debate. In the Dutch House of Commons, interruptions of speeches are made frequently and are recorded in the proceedings together with answers to interruptions. This may cause that governing parties also use more negative terms (e.g., in their answers to negatively phrased interruptions). An interesting dataset to test the findings of Hirst et al. (this volume) are the Danish parliamentary proceedings. Denmark has a minority cabinet since the early nineties. It could be interesting to see how the dynamic between government and parliament in this situation is reflected in the language of attack and defense.

References

Banea, C., R. Mihalcea, and J. Wiebe. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. *LREC 2008*.

Chesley, P., B. Vincent, L. Xu, and R. Srihari. 2006. *Using Verbs and Adjectives to Automatically Classify Blog Sentiment*. *AAAI Spring Symposium Technical Report SS-06-03*.

Compact Oxford English Dictionary: opinion. (n.d.). (O. U. Press, Producer) Retrieved 06 05, 2009 from Compact Oxford English Dictionary: http://www.askoxford.com:80/concise_oed/opinion

Ding, X., and B. Liu. 2007. The utility of linguistic rules in opinion mining. *Proceedings of the 30th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, pp. 811–812. Amsterdam: ACM, New York, NY.

Edens, J., M. Liem., T. Mensink, R. Weve, and L. van Zande. 2006. *Measuring Politics*. University of Amsterdam, Amsterdam.

Esuli, A. and F. Sebastiani. 2006. Determining term subjectivity and term orientation for opinion mining. *Proceedings of the Eleventh Conference on European Chapter of the Association for Computational Linguistics*. Trento, Italy: European Chapter Meeting of the ACL. Association for Computational Linguistics., pp. 193–200.

Furuse, O., N. Hiroshima, S. Yamada, and R. Kataoka. 2007. Opinion sentence search engine on open-domain blog. *Proceedings of the 20th International Joint Conference of Artificial Intelligence (IJCAI2007)*.

Jijkoun, V. and K. Hofmann. 2009. Generating a non-English subjectivity lexicon: Relations that matter. *Second Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*.

Kamps, J. and M. Marx. 2001. Words with attitude. *1st International WordNet Conference*, pp. 332–341.

Kim, S.-M. and E. Hovy. 2004. Determining the sentiment of opinions. *Proceedings of COLING-04*, pp. 1367–1373. Geneva, Switzerland.

Kim, S.-M. and E. H. Hovy. 2005. Automatic detection of opinion bearing words and sentences. *Second International Joint Conference on Natural Language Processing*.

Ku, L., Y. Liang, and H. Chen. 2006 Tagging heterogeneous evaluation corpora for opinionated tasks. *LREC 2006*.

Liu, B. 2007. *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*. Heidelberg: Springer.

Manning, C., P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511809071

Marx, M., N. Aders and A. Schuth. 2010. Digital sustainable publication of legacy parliamentery proceedings. In *Proceedings dg.o 2010*.

Marx, M. and A. Schuth. 2010. DutchParl. A corpus of parliamentary proceedings in Dutch. In *Proceedings LREC 2010*. pp. 3670–3677.

McKeown, K. and V. Hatzivassiloglou. 1997. Predicting the semantic orientation of adjectives. *Proceedings of the 35th annual meeting of ACL*.

Mishne, G. 2005. Experiments with mood classification in blog posts. *1st Workshop on Stylistic Analysis Of Text For Information Access*.

Mullen, T. and R. Malouf. 2006. A preliminary investigation into sentiment analysis of informal political discourse. *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 159–162.

Osgood, C., G. Suci, and P. Tannenbaum. 1957. *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Osman, D. and J. Yearwood. 2007. Opinion search in web logs. *Proceedings of the Eighteenth Conference on Australasian Database, 63*. Ballarat, Victoria, Australia.

Pang, B. and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2 (1-2), pp. 1–135.

Riloff, E. and J. Wiebe. 2003. Learning extraction patterns for subjective expressions. *Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, pp. 105–112.

TreeTagger 3.2. (n.d.). From TreeTagger 3.2: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Turney, P. 2001. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association For Computational Linguistics* pp. 417–424. Philadelphia, Pennsylvania: Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ.

Turney, P. and M. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4), pp. 315–346. DOI: 10.1145/944012.944013

Tweede Kamer: Plenaire vergaderingen. (n.d.). Retrieved 06 03, 2009 from Tweede Kamer der Staten Generaal: http://www.tweedekamer.nl/vergaderingen/plenaire_vergaderingen/index.jsp

Van Dale online dictionary: mening. (n.d.). (V. Dale, Producer) Retrieved 06 05, 2009 from Mening: http://www.vandale.nl/vandale/opzoeken/woordenboek/?zoekwoord=mening

WekaWiki: Primer. (n.d.). Retrieved 06 10, 2009 from Weka-Machine Learning Software in Java: http://weka.wiki.sourceforge.net/Primer

Wiebe, J. M., R. F. Bruce, and T. P. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 246–253. College Park, Maryland: Association for Computational Linguistics. DOI: 10.3115/1034678.1034721

Wiebe, J. and E. Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Computational Linguistics and Intelligent Text Processing* Vol. 3406/2005, pp. 486–497. Heidelberg: Springer.DOI: 10.1007/978-3-540-30586-6_53

Wilson, T., D. Pierce., and J. Wiebe. 2003. Identifying opinionated sentences. *Proceedings of the 2003 Conference of the North American Chapter of the Association For Computational Linguistics on Human Language Technology:Demonstrations. 4*, pp. 33–34. Edmonton, Canada: North American Chapter of the Association for Computational Linguistics.

Yu, H. and V. Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pp. 129–136. DOI: 10.3115/1119355.1119372

Appendix

```
<spreker pagina="" anker="252" naam="Van Raak" partij="SP" soort="Kamerlid"
        geslacht="man">
   <p gs:subjective="false">Voorzitter. 11 commissarissen van de Koningin hebben 35
      duurbetaalde bijbanen. Blijkbaar hebben ze niks anders te doen. Ik heb het heel
      druk, net als de minister. De minister heeft geen duurbetaalde bijbanen, maar die
      commissarissen wel. Waarom eigenlijk? Waarom hebben zij het niet net zo druk als
      deze minister? Wij moeten ze aan het werk zetten.</p>
   <p gs:subjective="true" gs:orientation="negative">Blijkbaar is een aantal deelnemers
      van het ABP, terugblikkend, toch niet zo tevreden over de heer Borghouts. Dat kan.
      Zij vragen nu advies aan Nout Wellink van DNB. Ik steun dat; ik hoop dat de minister
      haar politiemensen, die ook ernstige bezwaren tegen deze benoeming hebben, steunt.</p>
   <p gs:subjective="true" gs:orientation="negative">Ik heb de minister gevraagd om
      een keuze te maken. Dat doet zij echter niet. Als zij geen keuze maakt, maakt
      zij toch een keuze. Ik heb gezegd: als wij de commissarissen serieus nemen, dan
      moeten wij ze ook aan het werk zetten. Dat betekent dus dat zij geen bijbanen kunnen
      hebben. Ik neem de minister serieus en zij heeft ook geen bijbanen. Zeggen wij
      daarentegen: ja, die commissarissen van de Koningin kunnen wel al die bijbanen
      hebben, dan is het CdK-schap geen serieuze baan.</p>
</spreker>
<spreker pagina="" anker="257" naam="Verbeet" partij="" soort="Voorzitter"
        geslacht="vrouw">
   <p>Dank u wel.</p>
</spreker>
<spreker pagina="" anker="260" naam="Van Raak" partij="SP" soort="Kamerlid"
        geslacht="man">
   <p gs:subjective="true" gs:orientation="negative">Ik vraag dus de minister om een eind
      te maken aan al die duurbetaalde bijbanen van commissarissen van de Koningin. Dat is
      namelijk de enige manier om een serieuze baan van het CdK-schap van te maken.</p>
</spreker>
<spreker pagina="" anker="263" naam="Ter Horst" partij="" soort="Minister" geslacht="">
   <p gs:subjective="false">Voorzitter. Ik neem Kamerleden ook buitengewoon serieus. Zij
      mogen ook bijbanen hebben. Ik weet dus niet of dat een goede redenering is. De
      redenering zou moeten zijn dat het CdK-schap een serieuze baan is die serieus
      wordt betaald, maar dat het nuttig kan zijn -- ook voor de provincie -- dat een
      commissaris van de Koningin bijbanen heeft. Provinciale staten oordelen over het
      nut, een mogelijke belangentegenstelling en een cumulatie van bijbanen.</p>
   <p gs:subjective="true" gs:orientation="positive">Voor een groot deel ben ik het eens
      met de heer Van Raak over het financiële aspect. Daarom heeft heb kabinet besloten
      om een regeling te treffen voor de neveninkomsten van de commissarissen van de
      Koningin.</p>
</spreker>
```

Figure 1. Example of the data format in XML annotated with the gold standard.

CHAPTER 7

The qualitative analysis of political documents

Jared J. Wesley
University of Alberta, Department of Political Science

Qualitative document analysis remains one of the most common, yet methodologically misunderstood, components of political science research. While analysts are accustomed to incorporating manifestos, speeches, and other documents as evidence in their studies, few approach the task with the same level of understanding and sophistication as when applying quantitative methods. Building bridges between the two traditions, this chapter suggests guidelines for the rigorous, qualitative study of political documents. The discussion includes a novel examination of materials from the Poltext Project collection – a compilation of documents from across the Canadian provinces. The paper concludes that, whether approaching their work from a quantitative or non-quantitative perspective, researchers must adhere to similar disciplinary standards if their findings are to be considered trustworthy contributions to political science.

Introduction

Seldom do textbooks or courses in political science methodology devote any attention to manual qualitative document analysis (QDA). Dominated by the discussion of interviews, focus groups, (quantitative) content analysis, experimentation, and field studies, few offer any treatment of QDA, whatsoever (but see Wesley 2011a, 2011b, 2011c). When it is discussed, the qualitative analysis of political texts is typically included under broad topics like "unobtrusive" or "archival research," or conflated with the coding of transcripts or field notes (George 2006: 135; Platt 2006a: 83). This lack of detailed methodological discussion is disconcerting, considering that most political scientists have, at some point, incorporated elements of textual interpretation as evidence in their research. To move beyond "armchair interpretation," however, such analyses must be guided by the same level of rigour as those in the quantitative fields

(Wesley 2011b). The following paper suggests a series of such guidelines, based on a comparison with traditional modes of quantitative content analysis, a review of recent QDA studies, an examination of similar approaches in other social science disciplines, and a novel study of documents drawn from the Poltext Project collection.

Three types of Qualitative Document Analysis (QDA)

A review of the literature reveals that political scientists tend to employ three main forms of qualitative inquiry when it comes to the study of documents. Each involves its practitioners in a search for a different set of themes, using conventional, directed, or summative tools.

The first and most common qualitative approach to political documents involves *rhetorical analysis*. Rhetoric is the art of persuasion; thus students of rhetoric in political documents must investigate what the author of a text is trying to achieve and the strategies employed to that end. In this vein, documents may be analyzed according to their style of writing (e.g., formal or informal), type of diction (e.g., jargonistic or folksy), tone (e.g., moderate or extreme), use of certain devices (e.g., metaphors or sarcasm), reference to keywords or formulaic phrases (e.g., patriotism or feminism), invocation of certain emotions (e.g., certainty or fear), use of icons or figures (e.g., photographs or charts), allusion to imagery or symbolism, or any other element of the author's delivery.

In this sense, the manifest content of the appeal is of less concern to rhetorical analysts than purpose for, and the means by, which the message is conveyed. In their effort to provide "A Rhetorical Perspective on the 1997 British Party Manifestos," for example, Craig Allen Smith and Kathy B. Smith (2000:458) examined how political parties used "the resources of language to negotiate a shared understanding" among members of the electorate. As they describe, "Any election campaign is a contest between a 'we' and one or more 'thems' in which each party attempts to maximize the 'we' relative to the 'thems'… Political parties try to accomplish this through the management of themes, visions, symbols, needs, preferences, and reasons."

To study these rhetorical strategies, Smith and Smith examined each major party's campaign literature, comparing their ideological content, style of writing, iconic layout, and aesthetic packaging. "The Conservative Manifesto is a startlingly impersonal document," they reported.

> It shows only two human faces, virtually identical pictures of a suited [Prime Minister] John Major (one indoors and smiling, the other outdoors with wind-tousled hair and a half-smile. All other Britons are silhouetted, shadowed, or masked from view by looking into a microscope, facing a computer screen, wearing surgical masks, or talking on cell phones. These seem not be persons at all but subjects whose personal qualities are unknown to their leaders.
>
> (Smith and Smith 2000: 462)

By contrast, the analysts report that Labour's manifesto is dominated by photos and collages, featuring leader Tony Blair and a diverse collection of Britons. According to their interpretation, "Labour's message is that Great Britain is a land of happy people sharing life… In stark contrast to the Conservatives, Labour's is a populist portfolio. If Labour is to be feared, there is little corroborating evidence in these pictures of happy families, commuters, and schoolchildren. Visual identification is an important feature of this manifesto" (Smith and Smith 2000: 463). As Smith and Smith demonstrate, the non-verbal aspects of a text are just as important in persuasion as verbal communication. Rhetorical document analysis encompasses both.

A second qualitative approach to the study of political texts involves *discourse analysis* (see Eleveld and Filardo Llamas, this volume). Instead of examining the mode of communication, the discourse analyst investigates the broader values, norms, ideologies, and other contextual factors embedded in a particular (set of) document(s). Whether from a critical, normative, or empirical perspective, the objective of discourse analysis is to uncover the ideational foundation underpinning a particular text. The "discourse," itself, consists of the language – the bounded range of acceptable terms and statements – in which the document was written. As Louse Phillips described in her examination of "Hegemony and Political Discourse: The Lasting Impact of Thatcherism" (1998: 851), "Discourses dictate what it is possible to say and not possible to say and thus constrain other forms of social action." In this sense, the basic premises of discourse analysis are two-fold: "concretely, to elucidate the role of discursive practices in securing discursive change and hence cultural change…; at a more general level, to demonstrate that discourse (language understood as a *social practice*) is an active agent in social, political and cultural change (rather than just reflecting change)" (Phillips 1998: 848) (see also Berman 2001; Frankenberg 2006: 450). This requires the analyst to investigate, empirically, how the discourse is developed and propagated through "*actual instances of language use*" (Phillips 1998: 852, emphasis in original).

Nigel Copsey's "Reflections on the ideological evolution of the British National Party [BNP]" (2007) constitute a prime example of discourse analysis. Examining

party manifestos, campaign literature, media interviews, and public speeches, he assessed the claim that the BNP has transformed its public discourse from one rooted in 1930s fascism to a more modern form of national-populism. To test this hypothesis, Copsey drew extensively on the language of British National Party leader, Nick Griffin, who characterized his own 2005 manifesto – *Rebuilding British Democracy* – as "the final and decisive ideological paradigm shift from post WW2 camouflaged neo-fascism to 21st Century popular nationalism" (2007: 63). Copsey noted Griffin's abandonment of extremist attacks aimed at specific groups within British society, in favor of a more tempered assault on the "totalitarian", "repressive" machinery of the British state (Copsey 2007: 73).

Rather than taking Griffin at his word, however, Copsey conducted a systematic analysis of the BNP leader's own statements. Examining Griffin's discourse on a deeper level, he extracts the ideational foundation of the party; "what lurks within is not a commitment to western-style liberal democracy, but a core vision that represents a fundamental negation of democratic liberalism. That the BNP appears to be non-totalitarian should not distract us from the fact that it remains committed to revolutionary rebirth" (Copsey 2007: 77). According to Copsey's discourse analysis of BNP literature,

> the party's new ideological position should be treated with caution. The point that needs to be made here is that its "popular nationalism" or "national-populism" constitutes a surface ideology that does not change the party's core convictions… The reality is that at the root of the party's ideological modernization is short-term political expediency: not so much a change of course as an opportune change of clothing…. It is not the transformation from fascism to national-populism but the recalibration and modernization of fascism itself. It once more testifies to the almost Darwinian ability of fascism to survive and adapt as an ideology.
>
> (Copsey 2007: 79–80)

The discourse approach is particularly suited to uncover such shifts in party ideologies.

Narrative analysis constitutes the third major form of qualitative document analysis employed by political scientists. Unlike rhetorical analysts (who examine the delivery of the message) or discourse analysts (whose focus is on the ideas behind the message), narrative analysts investigate the content, origins, evolution, and impact of the message as a "story" about political life. Not all political documents contain narratives, of course. Those that do often present stories in the form of (un)official histories, myths, legends, folk tales, or personal accounts about the author or (members of) her political community. These narratives frequently tie the past to the present and future, speak of political transformations, and identify specific heroes, villains, and plotlines. These "storied" messages can be studied

from a variety of perspectives. Some analysts focus on the historical or political contexts in which the stories are told, for instance, while others examine the role of the stories in the lives of individual actors or their impact on the community, as a whole. In this sense, narrative analysts are not primarily concerned with the factual accuracy of the stories they uncover; rather, their concern is with the ways in which these stories serve as interpretive lenses through which the authors represent themselves and others.

In addition to their rhetorical analysis, Smith and Smith (2000) examined the narratives found in British manifestos, comparing the way Conservatives, Labourites, and Liberal Democrats portrayed the country's history during the 1997 election campaign. "No party waxed nostalgic. No party invoked heroes. No party relied on sacred national texts. And no party mentioned history prior to 1979," the analysts reported (Smith and Smith 2000: 465). "Instead, they recount the recent history that created the disputed state of affairs and the challenges of the future" (Smith and Smith 2000: 464). As the governing party, the Conservatives described the country's prosperity as the product of their record of "unparalleled success" and neo-liberal reforms, for instance. "The key to the future," according to Smith and Smith's interpretation of the Conservative manifesto, "is to do more of the same without knuckling under to those who would turn back" (Smith and Smith 2000: 464). The two opposition parties offered a different set of narratives. The Liberal Democrats lived "in a different present," according to the analysts, "one as gray as the landscape of a Welsh slate mine" (Smith and Smith 2000: 464). Smith and Smith support this interpretation by quoting directly from the Lib-Dem manifesto, which suggested "Eighteen years of Conservative government have left our society divided, our public services run down, our sense of community fractured, and our economy under-performing" (Smith and Smith 2000: 464). By contrast, Tony Blair's Labour platform was "more pragmatic than idealistic. 'I believe in Britain,' he says to begin the manifesto. 'It is a great country with a great history. The British people are a great people. But,' he continues meaningfully, 'I believe Britain can and must be better'" (Smith and Smith 2000: 464). Interestingly, according to the analysts, Blair's vision for the future marked a "sharp break form Labour's historic narrative" (Smith and Smith 2000: 465). Quoting Blair, they noted the Labour leader's

> "aim to put behind us the bitter political struggles of left and right that have torn our country apart for too many decades. Many of these conflicts have no relevance whatsoever to the modern world – public versus private, bosses versus workers, middle class versus working class. It is time for this country to move on and move forward. We are proud of our history, proud of what we have achieved – but we must learn from our history, not be chained to it" (p. 2).
>
> (Smith and Smith 2000: 465)

The narrative approach enabled the identification of the three different stories, narratives, from the three political parties.

This list of qualitative approaches is by no means exhaustive. Political documents are ripe for a variety of other inquiries. Philosophers may find them useful sources for hermeneutical examination, political scientists for party positioning, linguists for semiotic analysis, historians for process tracing, and critical analysts for deconstruction. Nor are rhetorical, discourse, and narrative analyses necessarily exclusive. Researchers will often combine these modes in a single study (see: Smith and Smith 2000). Furthermore, as other contributors to this volume demonstrate, concepts like rhetoric (Dahlberg and Sahlberg, Gryc and Moilanen, Leeuwen, Maks), discourse (Boyd, Montesano Montessori), and narrative may be analyzed quantitatively. Rather, the intent of the foregoing discussion is to highlight three of the most common approaches to the study of political documents, as a means of identifying the sorts of themes uncovered during the qualitative coding process. The main concern of this chapter is to ensure that the uncovering of themes through a qualitative process meets certain standards of empirical rigour. After a discussion of the ontological foundations of the quantitative and qualitative traditions, this chapter establishes a set of guidelines for the trustworthy practice of qualitative textual analysis.

Building bridges

Three ontological perspectives

The following discussion proceeds under the assumption that the qualitative and quantitative approaches to textual analysis are commensurable under the same, general standards of empirical inquiry. While distinct in many ways, each tradition contributes its own set of tools for the overall "toolkit" of the twenty-first century political scientist (Wesley 2011c). This view is only one of three general perspectives on the relationship between quantitative and qualitative methods (Bryman 2001: 276; Corbetta 2003: 50). While ideal types, in that no researcher is likely to adhere entirely or permanently to one set of beliefs, the distinctions are informative for document analysts and others who engage in research on either side of the quantitative-qualitative divide.

The first perspective holds that the quantitative and qualitative traditions are so ontologically distinct as to be incommensurable. Scholars in this "*purist*" school believe in a hard-and-fast connection between quantitative methods and the tenets

of positivism,[1] on one hand, and qualitative methods and interpretivism, on the other. According to this perspective, quantitative positivists believe in the principles of inherency and verifiability, which puts them at odds with the belief among qualitative relativists that all reality is socially constructed. As Manheim et al. (2002: 318) describe, "Some quantitatively oriented scholars regard at least some qualitative work as so dependent on the perceptions of the individual researcher and so focused on specific cases as to be unverifiable and essentially useless. In contrast, some qualitatively oriented scholars judge quantitative methods to be so incomplete in their representation of reality as to be empirically misleading…" In this environment, researchers toil in opposing camps – either parallel, but separate, in their pursuit of knowledge, or actively seeking to undermine the other.

Political science has not been immune to the above tensions; thankfully, however, "Most empirical researchers work primarily with either qualitative or quantitative methods but can see value in the other approach" (Manheim, Rich, and Willnat 2002: 318). This second perspective is embodied in the work of King, Keohane and Verba (1993), and holds that both quantitative and qualitative research methods are commensurable under the positivist approach to social life. In their words, "the differences between the quantitative and qualitative traditions are only stylistic and are methodologically and substantively unimportant. All good research can be understood – indeed is best understood – to derive from the same underlying logic of inference. Both quantitative and qualitative research can be systematic and scientific," provided each submits to "the rules of scientific inference – rules that are sometimes more clearly stated in the style of quantitative research" (1993: 4–5, 6). Critics of the "KKV" approach accuse the authors of developing a "quantitative template for qualitative research" – a premise that presupposes the superiority of the former over the latter (Brady, Collier, and Seawright 2004: 3). For this reason, qualitative purists have anointed King et al. as headmasters of the "quantitative imperialist" school, imposing positivist concepts like hypothesis-testing and inter-subjectivity on an unwilling qualitative community. In fairness to King et al., their aim was to bridge the "quantitative-systematic-generalizing" / "qualitative-humanistic-discursive" divide (King, Keohane, and Verba 1993: 4). Less pejoratively, then, one may refer to theirs as the "*neo-positivist*" perspective.

---

1. For the purposes of this discussion, "positivism" is defined by the following three tenets: (1) scientific methods (i.e., the testing of hypotheses derived from pre-existing theories) may be applied to the study of social life; (2) knowledge is only generated through observation (empiricism); and (3) facts and values are distinct, thus making objective inquiry possible (Snape and Spencer 2006).

A third perspective takes a middling view of the relationship between quantitative and qualitative methods. Developed most coherently in a volume edited by Brady, Collier and Seawright (2004), the "*dualist*" school promotes the coexistence of quantitative and qualitative traditions within a broad social scientific enterprise. Unlike "purists," "dualists" see value in collaboration between quantitative and qualitative researchers, and an important element of interdependence in their relationship. Compared to "neo-positivism," the "dualist" school sees strengths and weaknesses in both approaches. As Brady et al. (2004: 10) put it,

> *In the social sciences, qualitative research is hard to do well. Quantitative research is also hard to do well. Each tradition can and should learn from the other*. One version of conventional wisdom holds that achieving analytic rigor is more difficult in qualitative than in quantitative research. Yet in quantitative research, making valid inferences about complex political processes on the basis of observational data is likewise extremely difficult. There are no quick and easy recipes for either qualitative or quantitative analysis. In the face of these shared challenges, the two traditions have developed distinctive and complementary tools (emphasis in original). (Brady et al. 2004: 10)

Instead of struggling for methodological supremacy, dualists implore all social scientists to "refine and develop the battery of techniques on offer, and above all to be as explicit as possible about the implications of the methodologies we employ…" (Laver 2001: 9).

While acknowledging that many readers view the world from the "purist" and "neo-positivist" perspectives, the following discussion proceeds along "dualist" lines. According to this view, social science is the systematic study of the social world; the definition of what constitutes "systematic" is contentious, a debate that is explored in greater detail below.

Two traditions of document analysis

As the paradigm governing most quantitative research, positivism elevates two key concepts as crucial elements of any legitimate study in political science: validity and reliability. The former term refers to the importance of ensuring that one's findings accurately represent the concepts under examination. In content analysis, for example, a valid conclusion about the "positive tone" of a particular document must incorporate evidence of the author's optimism, cheerfulness, sanguinity, buoyancy, exuberance, or other senses of approbation. From a quantitative perspective, this evidence is often derived by counting the number of "positive" references, be they measured in terms of keyword mentions, phrases, sentences, paragraphs, or other units of analysis. Reliability, on the other hand, refers to the

consistency of a particular measurement – the extent to which a particular assessment would yield identical results if repeated under the same conditions. Content analysts typically measure this consistency through inter-coder reliability testing, a process in which the analyst's measurements are checked against those of an independent researcher.

As standards of academic rigour, both validity and reliability are rooted in the assumption that the information contained in documents is inherent – that the evidence embedded in the text is objectively identifiable. Armed with a list of pre-defined variables, the content analyst's task is to "mine" the documents in search of specific bits of data. This information is then analyzed, statistically, to discern important patterns existing within and between the documents.

While requiring a similar level of precision, the qualitative approach differs from quantitative content analysis in important ways. Rather than viewing data as inherent to the documents, themselves, most QDA researchers reject the notion of inter-subjectivity. From this perspective, "*the meanings invoked by texts need not be shared*" in a direct sense (Krippendorff 2004: 22–23, *emphasis in original*) (see also: Morse and Richards 2002: 125). To many QDA researchers, their particular interpretation of the text is just one of many possible "readings," thus imposing a different set of methodological burdens on them as they seek to convince their readers of the persuasiveness of their analyses (Gerring 1998: 298; Laitin 1986: 13).

As Manheim et al. (2002: 317) point out,

> Quantitative researchers are usually able to employ some well-established rules of analysis in deciding what is valid evidence for or against their theory. These include such tools as measures of statistical significance and statistical tests of validity, as well as formal logic. Qualitative researchers generally lack this type of commonly agreed to and 'objective' tool. Rather, they must rely on their ability to present a clear description, offer a convincing analysis, and make a strong argument for their interpretation to establish the value of their conclusions. Advocates of qualitative methods argue that this is an inevitable result of seeking to deal with the richness of complex realities rather than abstracting artificially constructed pieces of those realities for quantitative analysis. Critics of their approach contend that the vagueness and situational nature of their standards of evidence make it difficult (if not impossible) to achieve scientific consensus and, therefore, to make progress through cumulative knowledge. (Manheim et al. 2002: 317)

One particularly stinging critique holds that the findings of most qualitative analyses tend to be "conjectural, non-verifiable, non-cumulative, 'meanings'… arrived at by sheer intuition and individual guesswork" (Cohen 1974: 5). In short, qualitative researchers are subject to the criticism that they leave their readers with little choice but to "trust" that their interpretations of the data are accurate and legitimate.

Trustworthiness in political science

To guard against these criticisms, disciplinary standards require all political scientists to adhere to certain "rules" when it comes to treating texts as data. In particular, both quantitative content analysts and qualitative document analysts must establish the legitimacy of their research by protecting its "trustworthiness" in the eyes of the broader academic community.

The notion of "trustworthiness" is borrowed from the seminal research of Egon Guba and Yvonna Lincoln (1985). According to their dualist view, while quantitative research tends to be conducted under the premises of positivism, and qualitative inquiry under the auspices of interpretivism, there is considerable middle ground upon which to build consensus over the norms of social scientific analysis. Whereas scholars working in the qualitative tradition tend to reject the objectivity embedded in concepts like validity and reliability, for instance, Guba and Lincoln found that they tended to value and impose a similar set of standards on their own work. Their book, *Naturalistic Inquiry*, served as a sort of Rosetta Stone for interpreting four such common norms.

First, according to Lincoln and Guba's advice, all document analysts must protect the *authenticity* – or "truth value" – of their research. An authentic analysis is one that offers a genuine interpretation of reality, or an accurate reading of a particular (set of) document(s). This is referred to as "measurement validity" in the quantitative-positivist tradition, and "credibility" in the qualitative-interpretivist tradition. For the latter, the objective is less to offer a "truthful" account of the information found in the document, than to provide a believable interpretation of the meanings found therein (Richerson and Boyd 2004: 410–411). The authenticity of a qualitative analysis, then, relies upon the subjective evaluation of the reader, as opposed to being based against some objective standard (Krippendorff 2004: 314).

Portability is a second concern for analysts dealing with political documents. To make a substantive contribution to knowledge, most social scientists concur that their inquiries must offer insights extending beyond the specific cases under study (Bryman 2004: 539). In quantitative-positivist terms, this is referred to as "external validity" – the generalizability of a particular analysis to broader questions about political life. Content analysts strive to convince their readers that their findings can be expanded to other documents, from other sources, times, or places, for instance. The term "transferability" is preferred among those conducting QDA, once again reflecting their reluctance to accept the inter-subjectivity of their interpretations. Rather than establishing the generalizability of their analyses through tests of statistical significance, for example, qualitative document analysts rely upon their readers to assess the broader applicability of the lessons drawn

from their findings. In this sense, the question of whether the results of a qualitative document analysis can be extended to another context must be answered – not by the *original* investigator – but by the student seeking to make the transfer (Lewis and Ritchie 2006: 145; Merriam 2002: 228–229).

Third, researchers studying political documents must be wary of the *precision* of their analyses. Discussed above, content analysts tend to assess this aspect of trustworthiness in terms of reliability, through inter-coder testing. While replicability is fundamental to the positivist approach, however, its relevance is more contentious in the interpretivist tradition. As a consequence, many qualitative document analysts use the term "dependability" to describe the precision of their research. This captures the belief that, provided the research is conducted in a transparent manner, readers may assess the accuracy of the findings by asking, "Would I have reached the same general conclusions, given the opportunity to read the same set of documents under similar conditions?" An affirmative answer would confirm the dependability of the analysis.

The fourth and final concern among document analysts surrounds the *impartiality* of their observations. Social science is premised on the capacity of its practitioners to produce relatively unprejudiced knowledge about the social world, through findings that are reflective of reality as opposed to their own pre-determined beliefs (Marshall and Rossman 1989: 147). In quantitative research, this means preserving the "objectivity" of the analysis. Because they are more likely to consider personal biases to be unavoidable, if unfortunate, factors in the research process (King, Keohane, and Verba 1993: 14–15; Merriam 2002: 5), qualitative document analysts tend to acknowledge (even embrace) the subjectivity of their interpretations. To remain impartial, they must achieve "confirmability" in their findings, ensuring that their conclusions are drawn from the evidence at hand, as opposed to the predispositions of the researcher. The results of a QDA study are confirmable if the inferences drawn are traceable to data contained in the documents, themselves, and if the preponderance of evidence corroborates those findings. This is the very essence of empirical inquiry.

The foregoing discussion suggests that, relative to quantitative content analysis, QDA tends to place a heavier burden on the *reader* of the study to assess its trustworthiness. Some argue that this places too little responsibility on the *researcher* to defend the merits of the analysis. This criticism is misplaced, for in order to convince their audience of the trustworthiness of their research, qualitative data analysts must take equal care to meet the following disciplinary expectations of their work.

Achieving trustworthiness in qualitative document analysis

According to Guba and Lincoln (1994) the four norms of trustworthiness can be assured by (1) being explicit about the process by which the evidence is interpreted, and by (2) providing access to one's data so that findings may be verified. While providing for a post-hoc verification of the authenticity, portability, precision, and impartiality of their analyses, these two general safeguards are not sufficient to assure the trustworthiness of analyses. As Morse et al. (2002: 14) argue, by paying attention to the end of the study rather than the conduct of the research itself, investigators risk missing serious errors until it is too late to correct them. According to their assessment,

> in the time since Guba and Lincoln developed their criteria for trustworthiness, there has been a tendency for qualitative researchers to focus on the tangible outcomes of the research (which can be cited at the end of a study) rather than demonstrating how verification strategies were used to shape and direct the research during its development. While strategies of trustworthiness may be useful in attempting to *evaluate* rigor, they do not in themselves *ensure* rigor. While standards are useful for *evaluating* relevance and utility, they do not in themselves *ensure* that the research will be relevant and useful.
>
> (Morse et al. 2002: 16, emphasis in original)

Heeding Morse et al.'s advice, the following discussion provides a series of *guidelines* for the conduct of qualitative document analysis which are more specific than the first key assurance of Guba and Lincoln. Their compilation may create a useful checklist for reviewers, but their greater value lies in the support they provide for ensuring the trustworthiness of document analysis, be it quantitative or qualitative. Four sets of concerns are outlined, including those dealing with (1) triangulation, (2) intense exposure and thick description, (3) audit trails and discrepant evidence, and (4) intra- and inter-coder testing.

Triangulation

When using any form of data, political scientists are wise to corroborate their findings using other types and sources of evidence. Document analysts are no different, in this respect (Boyatzis 1998: xiii). Whether combining their findings with interviews, focus groups, or other research strategies, or conducting a "mixed-methods" form of research involving both quantitative and qualitative forms of textual analysis, researchers using political documents as their primary sources of evidence must substantiate their findings with some form of external

support (Tashakkori and Teddlie 2003: x). For qualitative document analysts, this "triangulation" may take several forms.

The first, and most common, involves "quantizing" one's findings. This means buttressing any subjective, qualitative interpretations of the latent elements of a text with more objective, quantitative analyses of its manifest content (Hesse-Biber and Leavy 2006: 326–330). References to the existence of a particular "theme" in a set of documents, for instance, may benefit from an indication of how many times a particular set of keywords appeared in the texts. Doing so bolsters (the reader's confidence in) the precision of the analysis.

John Gerring applied this technique in his study of *Party Ideologies in America* (1998). Confronted with the choice between quantitative content analysis and qualitative document analysis, Gerring opted for the latter. "To make claims about party ideologies," he argued, "one must involve oneself in the meat and gristle of political life, which is to say in language. Language connotes the raw data of most studies of how people think about politics, for it is through language that politics is experienced" (Gerring 1998: 298). In this vein, Gerring (1998: 297) suggested, "it would be unrealistic to expect content analysis to bear the entire burden of analysis on a subject as vast and complex as party ideology. To begin with, one would be forced to scale back the quantity of evidence in a fairly drastic fashion… Second, and perhaps more significantly, content analysis is somewhat less scientific than it appears. Since the meaning of terms is not static or univocal, words do not fall automatically within content analysis categories."

In search of this "language," Gerring turned to American party platforms, dating back to 1828. There, he found distinct rhetorical patterns, such that Whig-Republicans spoke in terms of "order versus anarchy" and "the state versus the individual" and the Democrats in terms of "liberty versus tyranny" and "the people versus the interests". To substantiate his interpretation, Gerring provided detailed paraphrasing and copious quotations from the party platforms. Yet he also bolstered this analysis with a content analysis of specific terms, phrases, and concepts used by the various parties. By including graphs depicting the differentiated use of words like "liberty" or "the people", over time, Gerring effectively quantified his qualitative findings.

A second method of triangulation involves "member-checking" – a familiar tool to those conducting field or focus group researchers. These analysts often verify the results of their observations with the subjects, themselves, as a means of verifying the authenticity of their findings. In document analysis, this means consulting the authors of the texts, to see if one's interpretations match their original motives or intent. It may not be possible to consult the author of a specific document, whether due to death, anonymity, location, or disinclination. Where

possible, however, member-checking should be viewed as a valuable, "continuous process during data analysis… [not simply] as verification of the overall results…" (Morse et al. 2002: 16).

There may be disagreement between the researcher and the author, of course; indeed, there is often a healthy tension within research conducted from the *emic* and *etic* perspectives. In this sense, an author's intent may not be conveyed effectively to his or her audience, in which case the researcher's interpretation may provide a more authentic account of the *reader's* view of a particular text. Moreover, given that QDA is conducted with the understanding that no two readers are likely to come to identical interpretations of a given text, some disagreement between the author and the researcher is to be expected. The objective of member-checking is to provide at least some safeguard as to the authenticity of the researcher's interpretation of the text. Any disjunction between those findings and the author's own interpretation should not be taken necessarily as a refutation of the former, but rather a point to be explored during the analysis. The analyst should be prepared to defend his or her interpretation as the more trustworthy account of the text.

Intense exposure and thick description

A second set of guidelines requires qualitative document analysts to immerse themselves in their texts and produce detailed accounts of their findings. Some people refer to this process as one of "soaking and poking," although the imagery belittles the amount of rigour involved (King, Keohane, and Verba 1993: 36–43; Putnam 1993: 12; Shively 1998: 17). Granted, like any researcher, document analysts ought to "marinate" in their data until no new, alternative interpretations appear to emerge. (This is often referred to as the "saturation point.") Yet, to offer a trustworthy and systematic account, this process must be analytical and empirical, not simply a matter of osmosis.

Among political scientists, the coding process, itself, is one of the least-examined elements of qualitative document analysis. As health researchers Hsiu-Fang Hsieh and Sarah E. Shannon (2005) describe, there are three general techniques of qualitative textual coding, the choice among which depends largely upon the research question at hand. The first type of "conventional" document analysis is employed in exploratory studies, where existing theories or data in the subject area are limited. Under these circumstances, analysts cannot rely on previous research as a guide, but must allow the themes to emerge from the texts, themselves.

In engaging in this conventional, inductive style of inquiry, analysts typically rely on a systematic three-stage coding process (Miles and Huberman 1994).

During the first "open coding" phase, the researcher selects a small sample of the documents for a preliminary, in-depth review. She makes general notes about the broad themes that characterize each document individually, and all texts collectively. These themes are knitted together during a second stage of "axial coding," in which all documents are consulted. Patterns are given specific labels, and certain passages are "tagged" as belonging to one or more categories (see: Boyatzis 1998: 31). A third stage of "selective coding" involves checking and rechecking these tags, ensuring that labels are applied properly and noting any discrepant evidence (see: Creswell 1998: 150–152; David and Sutton 2004: 203–212; Morse and Richards 2002: 111–128; Neuman and Robson 2007: 337–342; Punch 2005: 199–204).

Consider a recent study of party platforms in the three Canadian Prairie Provinces of Alberta, Saskatchewan, and Manitoba (Wesley 2011b). Like Gerring, the researcher was motivated by the desire to uncover patterns in the rhetoric of dominant parties in each province. His specific intent was to discern whether the different political cultures found across the region were connected, in some way, to the elite level discourse taking place during election campaigns. Was Alberta's conservative political culture associated with the rhetoric of its dominant politicians? Did Saskatchewan politicians campaign with a left-wing accent, matching the social democratic ethos of the province? And was Manitoba's political culture of moderation connected to the campaign messages of its leading parties?

Armed with these initial questions and hypotheses, the analyst entered the inquiry by assuming that no such patterns existed, seeking evidence to support their presence and reject the null hypothesis. In search of these themes, he collected and analyzed over eight hundred pieces of campaign literature, dating back to 1932.[2] During the open-coding stage, the analyst detected certain broad-ranging themes. In Alberta, the discourse appeared to revolve around notions of liberty and anti-conformity, whereas parties in Saskatchewan tended to campaign on the importance of solidarity and community, and Manitoba on tolerance and modesty. These observations were recorded in the form of memos, which were used to direct the second phase of axial coding.

During this second stage, key passages were highlighted as belonging under the broad categories identified during the first read-through. Statements in Alberta were tagged as belonging under the category of "freedom," including subthemes like populism, individualism, and provincial autonomy. In Saskatchewan, axial coding classified certain statements as being evidence of the province's

2. Many of these documents are available online, as part of the Poltext Project collection.

"security"-based discourse, including references to collectivism, strong government, and polarization. Manitoba party platforms were coded for evidence of "moderation" (incrementalism, pragmatism, and a loose form of partisanship). Throughout this process, the analyst was on constant guard for discrepant evidence, including themes that may have appeared across provincial borders.

A third, systematic pass through the documents involved significant reflection and revision. Some passages were reassigned to different categories, for instance, and numerous cases were highlighted that challenged the tidiness of the earlier analysis. Some Alberta politicians made use of social democratic rhetoric, for instance, while parties in Saskatchewan and Manitoba occasionally invoked conservative terminology during campaigns. This discrepant evidence was recorded, reported, and addressed in the final report (see below). Through this three-stage process, the analyst was able to systematically identify several themes, refine their content, and support their existence with evidence drawn from the documents, themselves.

Hsieh and Shannon label a second, more structured technique as the "directed" mode of qualitative textual analysis. In this approach, the researcher is able to rely on previous research for direction when interpreting the content of various documents. This prior knowledge may be used to conceptualize the research question, or to refine the coding scheme prior to analysis. In other words, "Using existing theory or prior research, researchers begin by identifying key concepts or variables as initial coding categories… Next, operational definitions for each category are determined using the theory" (Hsieh and Shannon 2005: 1281). Armed with this scheme, the researcher then reviews the documents, coding separate passages as matching various pre-determined themes and recording any discrepant evidence. Depending on the fitness of the data to the coding scheme, categories may require deletion or revision, and additional themes may be identified.

Consider Michelle Weinroth's (2004) qualitative study of the Canadian Liberal Party's "marketing strategy" in its "Anti-Deficit Campaign" during the mid-1990s. Her approach was a combination of rhetorical, discourse, and narrative analysis, as she sought to account for the success of the Liberal Party in maintaining its popularity amidst massive cutbacks in social services. "The questions that I pose," she wrote, "reflect a line of inquiry addressed by several critics of neo-liberalism" (Weinroth 2004: 46). Many of these researchers focused on the *economic* and *fiscal* discourse generated by a "nexus" of corporate, academic, media, and political elites. Using this literature to contextualize and conceptualize her own study, Weinroth (2004: 46) expanded upon earlier findings through her own unique interpretation of the data:

> While my own explanatory model shares much with these studies, its distinctiveness lies in its conceptual framework, for it stresses, unlike these other treatises, that the Liberal campaign deployed a symbolic language of *nationalism* in fiscal form, and that the mystique and persuasive power of such an ideology resides in its archetypally dramatic pattern…My essay turns its attention to the theatrical and nationalist dimensions of such propaganda in an effort to show the workings of a key episode in Canadian consensus-building (emphasis added).
>
> (Weinroth 2004: 46)

A third qualitative technique identified by Hsieh and Shannon is known as "summative content analysis." Instead of examining the texts in a more holistic sense, the researcher explores the different meanings attached to specific concepts found therein. After searching the documents for manifest content – that is, the appearance of the term(s) under study – the investigator probes the different contexts in which the particular words were employed. Did certain types of authors use the term(s) in similar ways, for example, or has the usage of the concept changed over time? After reviewing the documents in search of responses to these types of questions, the researcher returns to the text to classify each appearance of the term(s) as belonging to one or several categories.

Roderick P. Hart and his colleagues used summative content analysis in their investigation of "The American People in Crisis" (2002). Drawing on the rhetoric contained in Congressional speeches surrounding the events of September 11th and the Clinton Impeachment proceedings, the researchers examined the use of the term "American people" during periods of national turmoil. Hart et al. found that politicians referred to the citizenry in a variety of different ways, which, upon further examination and using previous literature as guidance, they divided into six main categories (the People's Time, the People's Situation, the People's Role, the People's Actions, the People's Qualities, and the People's Opponents).

When applying any of the three techniques described by Hsieh and Shannon, researchers must be meticulous in reporting the results of their analysis, and the evidence upon which the interpretations were based. Known famously as "thick description," Gerring refers to this process as grounding one's findings

> in copious quotations from the principals. At times, this may seem laborious. However the inclusion of actual language in a rhetoric-centred study should be seen as equivalent to the inclusion of raw data in a qualitative study; both allow the reader to evaluate the evidence without relying entirely on the author's own authority. It also provides a depth otherwise lacking in discussions of abstract concepts and content-analysis statistics. (Gerring 1998: 298)

This begs the question, however: "How much data is enough to substantiate one's findings?" Without enough supporting evidence, a qualitative document analysis amounts to little more than "armchair interpretation". Even with proper citations, too much paraphrasing may lead readers to question the authenticity and impartiality of the study. Conversely, "How much data is too much?" Without the researcher's own voice, a study consisting of too many direct quotations amounts to transcription, not inquiry (Morse and Richards 2002: 188). Striking a balance between evidence and analysis is especially challenging for QDA researchers, in this regard (Platt 2006: 111–112).

While no disciplinary convention exists, as "a safe rule of thumb," Berg (2004: 270) recommends including at least three pieces of corroborating evidence for each major interpretation. These may come in the form of direct quotations or detailed paraphrases, and – depending upon the researcher's style and chosen venue – can be incorporated in-text or by way of footnote. This places the onus on the analyst to support his or her interpretation with adequate evidence, while providing the reader with the assurance that the findings are not derived arbitrarily.

Audit trails and discrepant evidence

In order to provide readers with the opportunity to assess the authenticity and precision of their analyses, researchers must also report the exact process through which they achieved their results. In quantitative analysis, this is most efficiently accomplished through the publication of the research instrument (questionnaire, coding manual, or other guides). With no standardized instrument, qualitative document analysis requires that its practitioners provide detailed accounts, not only of their findings, but of the *process* by which they reached their conclusions (Platt 2006: 116). This entails creating an "audit trail" and reporting any discrepant evidence that may challenge their interpretations (Altheide 1996: 25–33).

As Holliday (2007: 7) suggests, most qualitative research involves making sense of the often "messy reality" of social life. Doing so requires the qualitative document analyst to make dozens of difficult and subjective decisions throughout the research process – choices about which similarities and differences constitute real "patterns" in the text; to what degree certain parallels constitute genuine "themes"; which titles should be used to identify these themes; which passages constitute solid "evidence"; how much discrepant evidence must exist to refute a particular set of findings; and many others. There are no inherently right or wrong answers to such questions; there are only stronger or weaker justifications of these choices. As a result, qualitative document analyst must be explicit in identifying and defending the various decisions they made throughout the research process.

Contrary to most methods textbooks, an audit trail is not constructed solely at the end of the study. While a post-hoc report may satisfy the needs of reviewers, documenting the development of a completed analysis does little to ensure that the study, itself, is conducted in a trustworthy fashion (Morse et al. 2002: 16). Rather, researchers must keep detailed records of their progress throughout the data gathering, analysis, and reporting stages. These notes are often kept in the form of memos or journals, and serve two objectives. For the benefit of the reader, they allow the analyst to more accurately report the outcome and rationale behind the various decisions made. Second, the process of chronicling, itself, serves a valuable purpose, as it ensures the analyst is aware of, and continuously seeking to justify, the many choices made throughout the inquiry.

Some of the most important decisions concern how to deal, and whether to report, discrepant evidence. By including only information that serves to confirm their interpretations of the text, qualitative data analysts often face criticism for offering analyses that are "too tidy" or "circular". On the latter, critics of some QDA studies cite the researchers for entering the analysis with pre-defined hypotheses that, in turn, determine what they "see" as significant (George 2006: 155). To avoid succumbing to this tendency, Becker (1998), Esterberg (2002: 175), and Berg (2004: 184) recommend employing the "null hypothesis trick," by which the analyst enters the inquiry assuming that no patterns exist; he or she must then assemble evidence from the documents to establish the existence of any themes. This helps shift the 'burden of proof' onto the researcher, and away from the reader.

Of course, qualitative document analysts need not "prove" the "truth" of their interpretations beyond all doubt. Most social scientists operate on a different standard of trustworthiness, requiring their peers to establish the persuasiveness of their findings against competing interpretations. In quantitative research, persuasiveness is often measured in terms of *probability* (e.g., statistical significance), whereas qualitative researchers often speak in terms of *plausibility*.[3] Each term connotes a distinct, but related, standard of legitimacy. As Richerson and Boyd (2004: 410–411) note, "plausibility arguments" have three features in common with more conventional hypotheses developed under the positivist paradigm:

3. As a method of scientific explanation, plausibility arguments are well established in both the natural and social sciences. Indeed, plausibility arguments underlie much of what we "know" about the physical and social world; they underpin many of the theories and laws developed by mathematicians, physicists, archaeologists, evolutionary biologists, anthropologists, and others. Some plausibility arguments are used in exploratory research, developing hypotheses to be tested empirically in the future. Others are untestable, at least given current knowledge, technology, theory, or conditions. (Consider theories surrounding the existence of the "quark," for instance.) In cases like this, plausibility arguments often resist "proof" – or even "testing" – in the formal sense.

> (1) a claim of deductive soundness, of in-principle logical sufficiency to explain a body of data; (2) sufficient support from the existing body of empirical data to suggest that it might actually be able to explain a body of data as well as or better than competing plausibility arguments; and (3) a program of research that might distinguish between the claims of competing plausibility arguments. The differences are that competing plausibility arguments (1) are seldom mutually exclusive, (2) can seldom be rejected by a single sharp experimental test (or small set of them), and (3) often end up being revised, limited in their generality or domain of applicability, or combined with competing arguments rather than being rejected. In other words, competing plausibility arguments are based on the claims that a different set of submodels is needed to achieve a given degree of realism and generality… or that a given model is correct as far as it goes, but applies with less generality, realism, or predictive power than its proponents claim.
>
> (Richerson and Boyd 2004: 410–411)

Thus, when developing their interpretations, qualitative document analysts need not feel pressure to "prove" their reading is the only accurate one. In fact, they are encouraged to report evidence that places reasonable bounds on their findings. An accomplished QDA researcher

> considers not just one inferential hypothesis when reading and rereading the original communication material, but also many alternatives to it. He systematically weighs the evidence available for and against each of these alternative inferences. Thus, the results of his analysis, if fully explicated, state not merely (1) the favored inference and the content 'evidence' for it, but also (2) alternative explanations of that content 'evidence,' (3) other content 'evidence' which may support alternative inferences, and (4) reasons for considering one inferential hypothesis more plausible than others. (George 2006: 155)

Doing so, and reporting the specific decisions in the audit trail, boosts the credibility of the analysis (Holliday 2007: 167–181).

These three sets of guidelines, triangulation, intense exposure and thick description, and audit trails and discrepant evidence, provide a more detailed account of the first key assurance of Guba and Lincoln, making the process of interpretation more explicit. Their second key assurance is equally important, though. Qualitative document analysts ought to provide reasonable access to their raw materials. This is not simply as a courtesy to reviewers, or to protect against charges of inauthenticity, imprecision, or partiality. It is also crucial to the advancement of knowledge, as it permits other researchers to conduct their own inquiries without having to undergo the same painstaking process of collecting the raw materials. While not always possible (due to resource constraints or concerns over copyright or confidentiality), ideally documents should be placed in the public domain. Given advances in digital scanning and optical character recognition,

it is becoming increasingly easier to post texts online, in electronic form. One such collection has been amassed under the auspices of the Poltext Project (see also Collette and Petry, this volume).

Funded by a grant from the Fonds québécois de la recherche sur la société et la culture, and housed at Université Laval in Quebec City (and online at www.poltext.org), the Poltext project collects and provides access to political documents drawn from across Canada and over time.[4] Amassed by a research team from across the country, the open-source collection is one of the largest of its kind in North America. It contains a growing assortment of party platforms, throne speeches, budget speeches, and a variety of other political documents at both the federal and provincial levels, dating back to the 1960s. As such, the Poltext collection serves as an unparalleled source of data on democratic life in Canada. Of note, access to provincial-level data is especially valuable to comparative researchers, both in Canada and beyond. The ten Canadian provinces constitute an underused series of laboratories for the comparative study of public policy, party ideology, political rhetoric, and many other areas of political science research.

The main advantage of the Poltext project lies in its provision of raw textual data for both quantitative content analysts and QDA researchers. Unlike similar databases, the information found in the Poltext collection does not come as pre-packaged data. While useful, one of the drawbacks to many other manifesto collections lies in the fact that their users must conform their research questions and methods to the data; the resulting inquiries amount to secondary data analysis, rather than primary research. Recent changes to the Comparative Manifesto Project Database have made its collection free-access, as well.

Finally, to guard against partiality, qualitative document analysts ought to investigate and report their personal biases. QDA may be considered a form of "unobtrusive" research, in that it does not directly involve human *subjects*. However, as the researchers are key *instruments* of qualitative research – filtering the raw documents through their own personal lenses in order to produce "data" – it is important to investigate possible sources of contamination (Merriam 2002: 5). Thus analysts must undertake a process of critical self-reflection before and during the inquiry, and disclose the results as part of the final report (Creswell 2003: 182).

4. The use of data from the project for publication purposes is subject to the mention of the following source: "Poltext project (www.poltext.org) Université Laval (Québec). The Poltext project is funded by a grant from the Fonds québécois de la recherche sur la société et la culture."

Conclusion

To reiterate, the foregoing discussion serves as a set of *guidelines* for the conduct of trustworthy qualitative document analysis. This is by no means an exhaustive list (see Wesley 2009, 2011a). Nor is it intended as a checklist for evaluating the authenticity, precision, portability, or impartiality of QDA studies. Morse et al. (2002: 16) are correct:

> Using standards for the purpose of post-hoc evaluation is to determine the extent to which the reviewers have confidence in the researcher's competence in conducting research following established norms. Rigor is supported by tangible evidence using audit trails, member checks, memos, and so forth. If the evaluation is positive, one assumes that the study was rigorous. We challenge this assumption and suggest that these processes have little to do with the actual attainment of reliability and validity. Contrary to current practices, rigor does not rely on special procedures external to the research process itself. (Morse et al. 2002: 16)

Often dismissed as simply reading, reviewing, or interpreting texts, trustworthy qualitative analysis of political documents requires as much rigour as any other methodology. Scholars who employ QDA will encounter this reality firsthand when presenting their work for review by quantitatively-minded audiences. Reviewers will often ask, "are we simply expected to *trust* you, that *your* interpretation of these texts is valid and reliable?" Or, "who is to say *your* reading of these materials is most accurate?" While it is true that "trusting" the evidence provided in QDA studies is comparable to "trusting" the data presented in regression tables and other forms of quantitative presentation – in that the reader will seldom refer back to the source material to replicate the analysis themselves – such a response is unlikely to persuade one's critics. Instead, QDA scholars must incorporate checks and balances into their analyses, to demonstrate how their 'reading' of the texts is as trustworthy as anyone else's.

The tremendous advances made in quantitative content analysis in recent decades have dwarfed the development of similar innovations in QDA. Yet, qualitative document analysis remains a core methodology in political science research. The foregoing discussion serves as a baseline for similar progress in qualitative document analysis, and encourages students to further explore ways of improving the accuracy, scope, reach, and acceptance of QDA in the broader political science community.

References

Altheide, D. L. 1996. *Qualitative Media Analysis*. Thousand Oaks: Sage.

Becker, H. S. 1998. *Tricks of the Trade: How to Think about Your Research While You're Doing It*. Chicago: University of Chicago Press. DOI: 10.7208/chicago/9780226040998.001.0001

Berg, B. L. 2004. *Qualitative Research Methods for the Social Sciences*. 5th Edition. Toronto: Pearson.

Berman, S. 2001. Ideas, norms, and culture in political analysis: Review article. *Comparative Politics* 33(2), pp. 231–250. DOI: 10.2307/422380

Boyatzis, R. E. 1998. *Transforming Qualitative Information: Thematic Analysis and Code Development*. Thousand Oaks: Sage Publications.

Brady, H. E., D. Collier and J. Seawright. 2004. Refocusing the discussion of methodology. In H. E. Brady and D. Collier (eds.) *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Boulder: Rowman & Littlefield, pp. 3–20.

Bryman, A. 2001. *Social Research Methods*. Toronto: Oxford University Press.

Bryman, A. 2004. *Social Research Methods*. 2nd Edition. Toronto: Oxford University Press.

Cohen, A. 1974. *Two-Dimensional Man: An Essay on the Anthropology of Power and Symbolism in Complex Society*. Berkeley: University of California Press.

Copsey, N. 2007. Changing course or changing clothes? Reflections on the ideological evolution of the British National Party 1999-2006. *Patterns of Prejudice* 41(1), pp. 61–82. DOI: 10.1080/00313220601118777

Corbetta, P. 2003. *Social Research: Theory, Methods and Techniques*. Thousand Oaks: Sage Publications.

Creswell, J. W. 1998. *Qualitative Inquiry and Research Design: Choosing among Five Traditions*. Thousand Oaks: Sage Publications.

Creswell, J. W. 2003. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Thousand Oaks: Sage Publications.

David, M. and C. D. Sutton. 2004. *Social Research: The Basics*. Thousand Oaks: Sage Publications.

Esterberg, K. G. 2002. *Qualitative Methods in Social Research*. Boston: McGraw Hill.

Frankenberg, G. 2006. Comparing constitutions: Ideas, ideals, and ideology—toward a layered narrative. *International Journal of Constitutional Law* 4(3), pp. 439–459. DOI: 10.1093/icon/mol012

George, A. L. 2006. Quantitative and qualitative approaches to content analysis. In J. Scott (ed.) *Documentary Research*. Volume 1. Thousand Oaks: Sage Publications. pp. 135–160.

Gerring, J. 1998. *Party Ideologies in America, 1828–1996*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139174992

Guba, E. G. and Y. S. Lincoln. 1985. *Naturalistic Inquiry*. Thousand Oaks: Sage Publications.

Hart, R. P., S. E. Jarvis and E. T. Lim. 2002. The American people in crisis: A content analysis. *Political Psychology* 23(3), pp. 417–437. DOI: 10.1111/0162-895X.00292

Hesse-Biber, S. N. and P. Leavy. 2006. *The Practice of Qualitative Research*. Thousand Oaks: Sage Publications.

Holliday, A. 2007. *Doing and Writing Qualitative Research*. 2nd Ed. Thousand Oaks: Sage Publications.

Hsieh, H. and S. E. Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative Health Research* 15, pp. 1277–1288.

King, G., R. Keohane and S. Verba. 1993. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

Krippendorff, K. 2004. *Content Analysis: An Introduction to Its Methodology*. 2nd Edition. Thousand Oaks: Sage Publications.

Laitin, D. D. 1986. *Hegemony and Culture: Politics and Religious Change among the Yoruba*. Chicago: University of Chicago Press.

Laver, M. 2001. Why should we estimate the policy position of political actors? In M. Laver (ed.) *Estimating the Policy Positions of Political Actors*. London: Routledge. pp. 1–3

Lewis, J. and J. Ritchie. 2006. Generalizing from qualitative research. In J. Ritchie and J. Lewis (eds.) *Qualitative Research Practice: A Guide for Social Science Students and Researchers*. Thousand Oaks: Sage Publications. pp. 263–286.

Manheim, J. B., R. C. Rich and L. Willnat (eds). 2002. *Empirical Political Analysis: Research Methods in Political Science*. 5th Ed. Toronto: Longman.

Marshall, C. and G. B. Rossman. 1989. *Designing Qualitative Research*. Thousand Oaks: Sage.

Merriam, S. B. (ed.) 2002. *Qualitative Research in Practice*. San Francisco: Jossey-Bass.

Miles, M. B. and A. M. Huberman. 1994. *Qualitative Data Analysis: A Sourcebook of New Methods*. Thousand Oaks: Sage Publications.

Morse, J. M., M. Barrett, M. Mayan, K. Olson and J. Spiers. 2002. Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods* 1(2), pp. 13–22.

Morse, J. M. and L. Richards. 2002. *Read Me First for a User's Guide to Qualitative Methods*. Thousand Oaks: Sage Publications.

Neuman, W. L. and K. Robson. 2007. *Basics of Social Research: Qualitative and Quantitative Approaches. Canadian Edition*. Toronto: Pearson.

Phillips, L. 1998. Hegemony and political discourse: The lasting impact of Thatcherism. *Sociology* 32(4), pp. 847–867. DOI: 10.1177/0038038598032004011

Platt, J. 2006a. Evidence and proof in documentary research: Part I, some specific problems of documentary research. In J. Scott (ed.) *Documentary Research*. Volume 1. Thousand Oaks: Sage Publications. p. 83.

Platt, J. 2006b. Evidence and proof in documentary research: Part II, some shared problems of documentary research. In *Documentary Research*, edited by J. Scott. Volume 1. Thousand Oaks: Sage Publications. p. 105.

Punch, K. F. 2005. *Introduction to Social Research: Quantitative and Qualitative Approaches*. 2nd Edition. Thousand Oaks: Sage Publications.

Putnam, R. 1993. *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton: Princeton University Press.

Richerson, P. J. and R. Boyd. 2004. *The Origin and Evolution of Cultures*. London: Oxford University Press.

Shively, W. Ph. 1998. *The Craft of Political Research*. 4th Ed. Upper Saddle River, NJ: Prentice Hall.

Smith, C. A. and K. B. Smith. 2000. A rhetorical perspective on the 1997 British party manifestos. *Political Communication* 17(4), pp. 457–473. DOI: 10.1080/10584600050179068

Snape, D. and L. Spencer. 2006. The foundations of qualitative research. In J. Ritchie and J. Lewis (eds.) *Qualitative Research Practice: A Guide for Social Science Students and Researchers*. Thousand Oaks: Sage Publications.

Tashakkori, A. and Ch. Teddlie (eds). 2003. *Handbook of Mixed Methods in Social and Behavioural Research*. Thousand Oaks: Sage Publications.

Weinroth, M. 2004. Rituals of rhetoric and nationhood: The liberal anti-deficit campaign, 1994–1998. *Journal of Canadian Studies* 38(2), pp. 44–79.

Wesley, J. J. 2009. In search of brokerage and responsibility: Party politics in Manitoba. *Canadian Journal of Political Science* 42(1), pp. 211–236. DOI: 10.1017/S0008423909090088

Wesley, J. J. 2011a. Analyzing qualitative data. In *Explorations: A Navigator s Guide to Research in Canadian Political Science*, edited by K. Archer and L. Youngman-Berdahl. 2nd ed. Toronto: Oxford University Press.

Wesley, J. J. 2011b. *Code Politics: Campaigns and Cultures on the Canadian Prairies*. Vancouver: UBC Press.

Wesley, J. J. 2011c. Observing the political world: Quantitative and qualitative approaches In K. Archer and L. Youngman-Berdahl (eds.) *Explorations: A Navigator's Guide to Research in Canadian Political Science*. 2nd Edition. Toronto: Oxford University Press.