

Commentaries

Friends With Text as Data Benefits: Assessing and Extending the Use of Automated Text Analysis in Political Science and Political Psychology

Martijn Schoonvelde^{*a}, Gijs Schumacher^b, Bert N. Bakker^c

[a] School of Politics and International Relations, University College Dublin, Dublin, Ireland. [b] Department of Political Science, University of Amsterdam, Amsterdam, the Netherlands. [c] Amsterdam School of Communication Research, University of Amsterdam, Amsterdam, the Netherlands.

Abstract

Applications of automated text analysis measuring topics, ideology, sentiment or even personality are booming in fields like political science and political psychology. These developments are to be applauded as they bring about novel insights about politics using new sources of (unstructured) data. However, a divide exists between work in both disciplines using text as data. In this paper we argue in favor of more integration across disciplinary boundaries, structuring our case around four key issues in the research process: (i) sampling text; (ii) authorship as meta data; (iii) pre-processing text; (iv) analyzing text. Along the way we demonstrate that an assessment of speaker characteristics may crucially depend on the text sources under study, and that the use of sentiment words correlates with estimates of policy positions, with implications for interpretation of the latter. As such, this paper contributes to a critical discussion about the merits of automated text analysis methods in political psychology and political science, with an eye towards advancing the considerable potential of text as data in the study of politics.

Keywords: automated text analysis, political psychology, political science, text as data, literature review

Journal of Social and Political Psychology, 2019, Vol. 7(1), 124–143, <https://doi.org/10.5964/jspp.v7i1.964>

Received: 2018-03-28. Accepted: 2018-10-23. Published (VoR): 2019-02-08.

Handling Editor: Inari Sakki, University of Eastern Finland, Kuopio, Finland

*Corresponding author at: University College Dublin, School of Politics and International Relations, Newman Building Belfield, Dublin 4, Ireland. E-mail: mschoonvelde@gmail.com



This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Political scientists and political psychologists are turning to text to study politics, as evidenced by a proliferation of studies using automated text analysis methods to explore, for example, policy positions and topics, speaker sentiment or even personality (Benoit & Laver, 2007; Grimmer, 2010; Slapin & Proksch, 2008; Young & Soroka, 2012). These developments are to be applauded as they bring about novel insights about policies and politics using new sources of unstructured data. However, a divide exists between researchers using text as data in political science on the one hand and (political) psychology on the other, with cross-disciplinary work the exception not the rule. Generally, political psychologists are more likely to apply supervised methods like dictionaries (which assume that underlying categories are known) to learn from text about stable characteristics of the author or speaker (e.g., personality type or linguistic styles). Political scientists on the other hand more often use unsupervised

methods like topic models or scaling models (which assume that underlying categories are unknown) to learn from text about topical content or policy positions.¹ Each approach is valuable, but as this paper demonstrates they could both benefit from better integration (for a similar argument for political science and political psychology more generally see [Druckman, Kuklinski, & Sigelman, 2009](#)). For example, political scientists could learn from political psychologists about how individual characteristics are reflected in stable language patterns among politicians, whereas political psychologists could learn from political scientists how the political context (e.g., the dynamics a political campaign or the intended audience of a speech) pressures these politicians into changing their language use.

To further advance the promise it holds, this paper provides a multidisciplinary assessment of crucial assumptions in a typical text as data project, highlighting differences between political scientists and political psychologists. Building on [Grimmer and Stewart \(2013\)](#) and [Wilkerson and Casas \(2017\)](#), this assessment is structured around four central steps in a typical text as data research design: (i) sampling text; (ii) authorship as meta data; (iii) preprocessing text; and (iv) analyzing text. Our discussion is intended to raise awareness of each of these issues as well as to provide practical suggestions on how to deal with them. Along the way we demonstrate that assessment of speaker characteristics may crucially depend on the text sources under study, and that the use of sentiment words correlates with estimates of policy positions, with implications for interpretation of the latter. Our discussion is by no means intended to disqualify published results. Rather we want to highlight the importance of considering each of these issues when starting a text as data project. In the next four sections, we discuss each issue in turn. We then summarize our discussion by offering a set of best practices and finish with some concluding thoughts.

Sampling Text

The first issue in every text as data project concerns sampling. What text sources should be used to build a corpus? And what text sources should not be used? And why? A key consideration is – of course – to identify the text source best able to capture the theoretical construct of interest. Among political psychologists, researchers interested in personality or leadership style generally consider interview responses to be the most valuable source of text. The argument goes that in interviews the language used by the interviewee is more natural than in most other settings, and therefore ideal for capturing personality and style ([Hermann, 2005](#); [Slatcher, Chung, Pennebaker, & Stone, 2007](#); [Winter, Hermann, Weintraub, & Walker, 1991](#)). Other sources of “spontaneous” text are used as well. For example, in their study of linguistic styles among four presidential candidates, [Slatcher and colleagues \(2007, p. 64\)](#) note: “Although the final drafts of verbal texts yield useful knowledge about a person, more accurate indicators of people’s individual differences are spontaneous speech samples across varied social contexts. Among politicians, examples of available speech samples include press conferences, public interviews, and debates.” Thus, it is argued that as long as the text is produced “spontaneously” it can be analyzed for speaker characteristics, regardless of whether it is a debate text, an interview, a press conference or something else.

Work in political science on the other hand has argued that it is preferable to use similar text sources, because otherwise model output may be biased ([Gemenis, 2013](#)). The reason for this is that specific words may be more common in one text source than another, which may depend on the intended audience or the (political) process of how the text came about (i.e., the “data-generating process”). For example, [De Lange and Van Erkel \(2013\)](#) compared election manifestos of parties and subsequent coalition agreements to study which parties were most influential during the coalition formation process. Using *Wordscores* to scale these texts on an underlying dimension,

their results showed that coalition agreements were more extreme than all parties involved. As argued by these authors it is highly unlikely that coalition parties will decide on such a coalition agreement. A more plausible explanation is that language use varies between coalition agreements and election manifestos which leads the *Word-scores* procedure – rather than picking up on ideological differences – to simply distinguish between election manifestos and coalition agreements. This example is not unique. Biases may even emerge if texts serve similar purposes but the way they came about is different. For example, political parties use election manifestos for various purposes: sometimes they are the result of an extended process of negotiation between several party factions and in other cases the election manifesto is the expression of a powerful party leader. These variations in party organization can lead to differences in language use in manifestos as well.

Turning back to the example of spontaneity as a criterion for analyzing text for speaker characteristics, it is informative to note that in their well-cited analysis of linguistic style of U.S. presidential and vice presidential candidates, [Slatcher and colleagues \(2007, p. 69\)](#) report considerable aggregate differences between text sources – which they all consider to be spontaneous – among speakers in three out of six linguistic measures in the 271 texts under study: “Candidates used language more like that of a depressed person in interviews compared to press conferences ($d = .92$) and town hall meetings ($d = 1.19$); they used language more like that of an older person in press conferences compared to interviews ($d = .84$) and debates ($d = 1.30$) and in town hall meetings compared to interviews ($d = .56$) and debates ($d = 1.12$); their language was less presidential in interviews compared to press conferences ($d = 1.01$) and town hall meetings ($d = .83$).” That is, among the same set of speakers (George W. Bush, John Kerry, Dick Cheney and John Edwards) aggregate language use in press conferences, town hall meetings and interviews, varies on multiple dimensions, which – from the perspective of political scientists – makes sense since they are targeted to different audiences. Additional analysis of their data (see [Figure 1](#)) confirms that these patterns also exist within individual speakers.ⁱⁱ In a comparison of text sources for which we have at least 10 observations per speaker, it appears that George W. Bush scores significantly lower on honesty, aging and presidentiality during town hall meetings than during press conferences, and significantly higher on cognitive complexity. John Kerry, on the other hand, speaks more like an older, depressed, and less presidential person in his press conferences than in his network interviews. Substantive conclusions about their language use would thus depend on the type of text source used. This indicates that individual differences and political context together impact language use.ⁱⁱⁱ

The preceding discussion is not intended to disqualify these published results but rather to highlight that both political scientists and political psychologists will need to put in careful work when constructing a corpus. We propose that if analysts have reason to believe that text sources are systematically different from each other that they account for these differences in their models. For example, political scientists have developed the structural topic model ([Roberts et al., 2014](#)) which allows for meta data (such as author, type of audience, occasion, etc.) to influence model results. In sum, when constructing a corpus, we propose analysts use similar text sources to the extent possible. When a corpus consists of multiple text sources, analysts should account for this in their models as meta data. We will turn to a particular application of using meta data next.

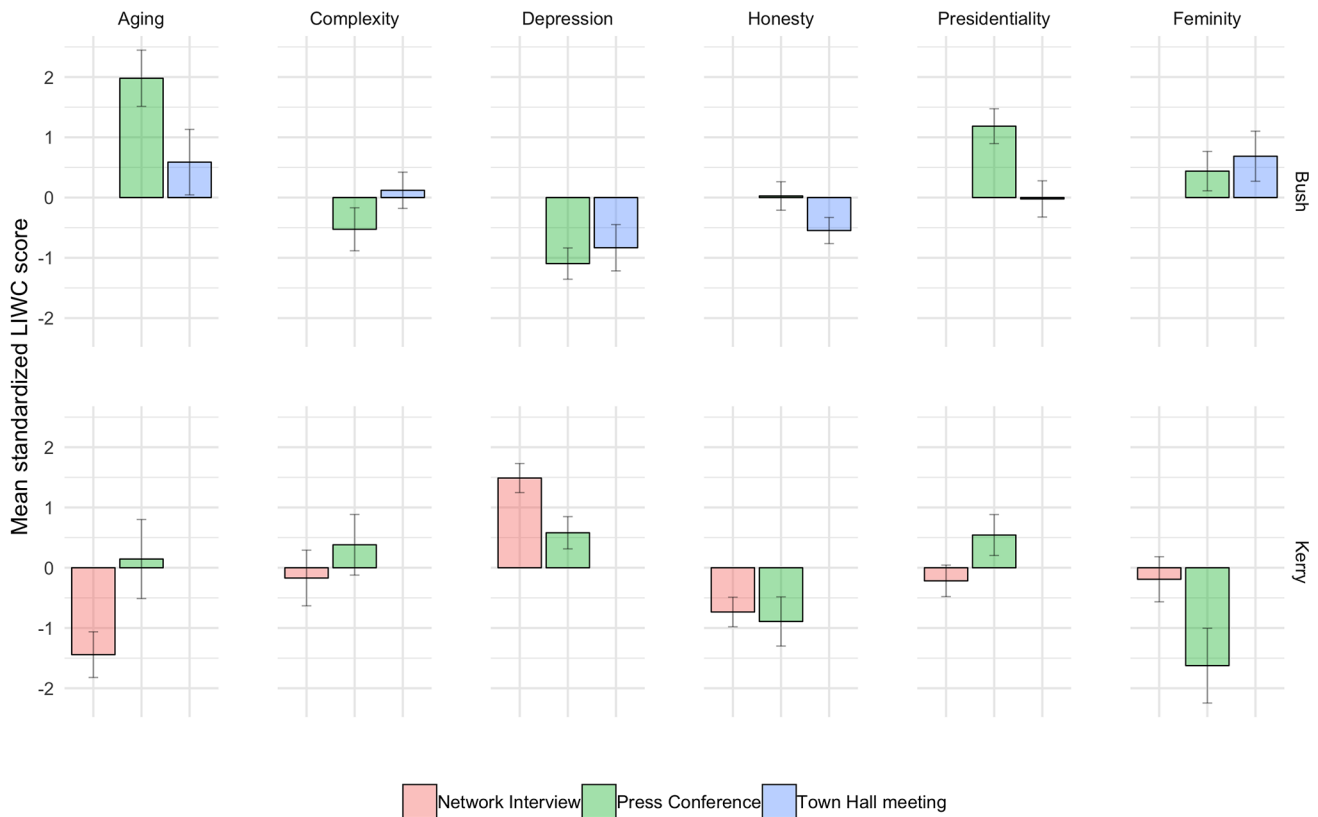


Figure 1. Linguistic style of George W. Bush and John Kerry on six linguistic style dimensions.

Note. This figure displays standardized LIWC scores for George W. Bush and John Kerry on six linguistic style dimensions (aging, complexity, depression, honesty, presidentiality, femininity) for various text sources for which we have at least 10 observations: network interviews (Kerry: $n = 44$), press conferences (Bush: $n = 57$; Kerry: $n = 21$), and town hall meetings (Bush: $n = 38$) (for more information, see Slatcher et al., 2007).

Authorship as Meta Data

Analysts interested in psychological constructs like personality may use politicians' speeches to measure such constructs "at-a-distance". This approach opens up many opportunities. For one, a direct approach of interviewing political elites is not always feasible since survey response rates among this group is generally low (Dietrich, Lasley, Mondak, Remmel, & Turner, 2012) and – important for those researchers interested in historical data – limited to those politicians who are still alive. The beauty of text analysis is that – once a text is archived – it can be studied, no matter what the time span. But of course, analysts face hurdles as well; there is no free lunch. Importantly, it is likely that not the politician but a speech writer wrote the text. Yet the impact this issue has on our ability to learn about the politicians' characteristics from these texts is far from clear. For example, a comparison between private recording of John F. Kennedy and his public speeches revealed no differences in leadership assessment (Renshon, 2009). Dille (2000) on the other hand found small but important differences in leadership style assessment between spontaneous and prepared remarks for George H. Bush and Ronald Reagan. However, if either president was involved in drafting the speech these differences disappear.

To understand their involvement, we should understand the role conception and incentives of speech writers. To this end we consulted guidebooks on becoming a speech writer and worked our way through several speech writer biographies. In terms of advice, a clear lesson we learned is that the speech should be an authentic and recognizable reflection of the “best possible version” of the speaker (Collins, 2012, p. 11). What is this best possible version? According to Collins (2012, p. 5), a speech performance is an artificial moment and your essential character will need to be drawn in “primary colours, sometimes in lurid colours, to make sure it is visible from the distant point in the audience”. Peggy Noonan, speech writer of Ronald Reagan, notes that “you have to find their sound” (Noonan, 1998, p. 101). “The way people speak usually reflects how they think. And so you must listen closely, not only so that the work you do sounds like them, but so it sounds like them thinking” (Noonan, 1998, p. 101). Following this advice, speech writers should write a speech in such a way that the personality of the speaker is visible to the audience. This way personality in speech may be slightly exaggerated, but probably not much.

But do actual speech writers stick to this advice? It depends. Obama speech writers – who referred to this speech writer Jon Favreau as his “mind reader” – had a lot of material to work with, but other speech writers were less fortunate. Jimmy Carter, for example, delivered few speeches about national issues and never met with his speech writers. This makes speech writing more difficult (Noonan, 1998), and perhaps in such cases personality assessment from speeches may be off. Consider the example of Barton Swaim (Swaim, 2015), a speech writer of former South Carolina Governor Mark Sanford. In contrast to the Obama-Favreau tandem, the working relationship between Swaim and Sanford was awful. Swaim had multiple speeches and op-eds sent back to him. He recalls: *“It was then that he [Nat, Sanford’s chief of staff] told me that everyone who worked for this governor had one goal. It wasn’t to please him with your superior work, because that would never happen. The goal was to take away any reason he might have to bitch at you. It was then too that Nat explained that my job wasn’t to write well; it was to write like the governor. I wasn’t hired to come up with brilliant phrases. I was hired to write what the governor would have written if he had had the time”* (Swaim, 2015, p. 9). The Swaim anecdote confirms that career incentives mobilize speech writers to write like their clients, not to write what they like. For Swaim his job became so awful that he commented: “Sometimes I felt no more attachment to the words I was writing than a dog has to its vomit” (Swaim, 2015, p. 6). For analysts on the other hand, it may strengthen confidence that one can learn about leader characteristics even when the text is written by a speech writer.

Based on our reading of these speech writer guidebooks, we are optimistic about the possibility of learning from speech writer text about leader characteristics. That being said, we encourage analysts to look into how the speech was produced. Does the politician have a speech writer at all? How many? Are they a long-running tandem or does the politician change speech writers often? And how is the politician involved in drafting the text? The analyst could also turn the search around and use supervised methods to predict whether politicians or speech writers wrote a text (see Airoidi, Fienberg, & Skinner, 2007) for an analysis of co-authorship of Ronald Reagan’s radio addresses). If the results of that effort prove to be unreliable, this may serve as evidence that speech writer text can indeed be used for learning about psychological constructs of the politician.^{iv} These are all pieces of information that are knowable but not often considered in political psychology nor in political science research. Just like with our previous discussion of sampling from different text sources, we consider authorship patterns a form of meta data which can tell us something about how a text came about. Rather than discarding such texts all together, their meta data should be included when the analyst builds a corpus.

Preprocessing Text

When analyzing political language, it is common to “preprocess” text in order to simplify the inputs to an analysis without altering its substantive conclusions. Common preprocessing steps encompass, for example, removal of numbers, punctuation and stop words, or word stemming. These preprocessing steps are typically presented as innocuous procedures, but in fact they may have non-trivial substantive consequences. For example, [Denny and Spirling \(2018\)](#) show how substantive conclusions from scaling methods and topic models may crucially depend on seemingly arbitrary decisions during preprocessing. To address this issue, [Denny and Spirling \(2018\)](#) propose that analysts first collect results from text analysis models with various different preprocessing steps. In a second step, the analyst evaluates if model results are robust to particular combinations of preprocessing steps. If model results are not sensitive to the applied preprocessing procedure, this increases confidence about their robustness. If model results vary with particular preprocessing steps, the analyst will need to report these dependencies.^v This is an important step forward in establishing robust results from unsupervised text as data models. However, it does not provide an explanation for *why* and *when* the results of unsupervised models depend on specific preprocessing steps. In this section, we argue that correlations between ideology, personality differences and linguistic habits could explain such patterns and we show evidence for this.

Work in psychology and linguistics reports that linguistic habits and the use of function words correlate with personality characteristics and policy positions ([Pennebaker, 2011](#)).^{vi} For example, introverted speakers prefer rich vocabulary ([Oberlander & Gill, 2006](#)), use more negations (e.g., [Pennebaker & King, 1999](#)) and fewer expressions and connectives ([Oberlander & Gill, 2006](#)). Neurotic speakers do not use rich vocabulary ([Oberlander & Gill, 2006](#)), are more likely to use first-person singular and to use words associated with negative emotions ([Pennebaker & King, 1999](#)). People high on openness to experience use more tentative words, such as ‘maybe’ or ‘perhaps’, and they use longer words ([Pennebaker & King, 1999](#)). People who score low on conscientiousness also use negations more often and are more likely to use negative emotion words ([Pennebaker & King, 1999](#)). At the same time there is an extensive literature that documents correlations between personality and ideology among citizens (e.g., [Bakker, 2017](#)) and political elites (e.g., [Caprara, Francescato, Mebane, Sorace, & Vecchione, 2010](#); [Dietrich et al., 2012](#)).

Given these correlations between personality characteristics, ideology and linguistic habits, common preprocessing steps may not be “ideologically neutral”. That is, they may affect some speakers more than others, leading to unreliable estimates from subsequent unsupervised models. This may in part depend on the amount of text data under study with the impact of preprocessing steps likely to be larger in smaller corpora.^{vii} These are two possible scenarios that we empirically evaluate using the EUSpeech dataset ([Schumacher, Schoonvelde, Dahiya, & De Vries, 2016](#); [Schumacher, Schoonvelde, Traber, Dahiya, & De Vries, 2016](#)). EUSpeech consists of all publicly available speeches from the main European institutions plus the IMF and the speeches of prime ministers – or president in the case of France – of 10 EU countries for the period after 1 January 2007: Czech Republic, France, Germany, Greece, Netherlands, Italy, Spain, United Kingdom, Poland and Portugal. From this dataset we select the English speeches from all the group leaders in the European Parliament as well as heads of government ($n = 3,301$), which are larger than 200 words. For each speech we count all stop words and divide that number by the total number of words in that speech to get at a proportion (the mean proportion of stop words across all speeches = 0.54). For this we used a stop word list from Quanteda ([Benoit & Nulty, 2016](#)) containing 174 words.^{viii} We collect similar statistics for 3 other common preprocessing steps in a typical text as data project:

stemming, the use of numbers and punctuation (see Denny & Spirling, 2018; Grimmer & Stewart, 2013; Wilkerson & Casas, 2017). Stemming concerns the algorithmic conversion of inflected forms of words into their root forms (for example, stemming the words “fish”, “fishing” and “fishes” converts them to “fish”, “fish” and “fish”). We use the Porter stemmer to get the number of unique tokens in a stemmed speech which we divide by the number of unique tokens in the same, unstemmed speech. The lower this proportion, the higher the impact of stemming (mean proportion of stemmed tokens across all speeches = 0.92). We also calculate the proportion of numbers to the total number of tokens in each unstemmed speech (mean proportion of numbers = 0.006). Furthermore, we calculate the number of punctuation tokens as a proportion of the total number of tokens (mean proportion of punctuation characters = 0.10). Punctuation tokens are dots, colons, semicolons, and so on.

To assess the degree to which these preprocessing steps correlate with political ideology, we collect two ideology measures from the Comparative Manifesto Group Database, which has systematically coded election manifestos to specific topics or positions on topics (Volkens, Lehmann, Matthieß, Merz, Regel, & Werner, 2016). We use the cultural progressive-conservative items^x and the economic left-right items^x from this database to calculate a progressive-conservative position and a left-right position for each speaker. Because the Manifesto Group contains data for each election, we use the score from their party’s most recent election manifesto as the speaker’s position.

We aggregate the proportion of numbers, punctuation, stop words and stemming for all 31 speakers in the corpus. Figure 2 shows the bivariate relationship between each four preprocessing steps and left-right and progressive-conservative ideology respectively using a Loess curve. These scatter plots reveal interesting patterns. For example, it appears that moving from the left to the center of the left-right dimension is positively correlated with the use of numbers: speakers in the center use on average more numbers than speakers on the left. Furthermore, moving from the progressive end to the center of the progressive-conservative dimension relates to an increase in the use of numbers as well. When comparing progressive and conservative speakers, we observe first a decrease and then a slight increase in the use of punctuation. We also find some variation across left-right and progressive-conservative speakers when it comes to using unique words: politicians on the extremes of these ideological scales tend to use slightly more unique words stems than politicians located in the ideological center (as evidenced by lower proportion and thus the higher impact of stemming for the latter). Tables B1 (stopwords), B2 (numbers), B3 (stemming) and B4 (punctuation) contain OLS regression results modeling at the speech level preprocessing scores as a function of left-right ideology, progressive-conservative ideology and their interaction, with fixed effects for countries (thus controlling for language specific differences). For each of the four preprocessing steps we find evidence that they are related with left-right and progressive conservative scores of the speakers. Taking the proportion for stop words as an example, the model estimates that Gabriele Zimmer and Lothar Bisky, the left-most speakers in the corpus (left-right score of -3.2) use about 10.5 punctuation characters per 100 words, and Nigel Farage (left-right score of +0.8) about 12 punctuation characters (who thus uses shorter sentences on average).

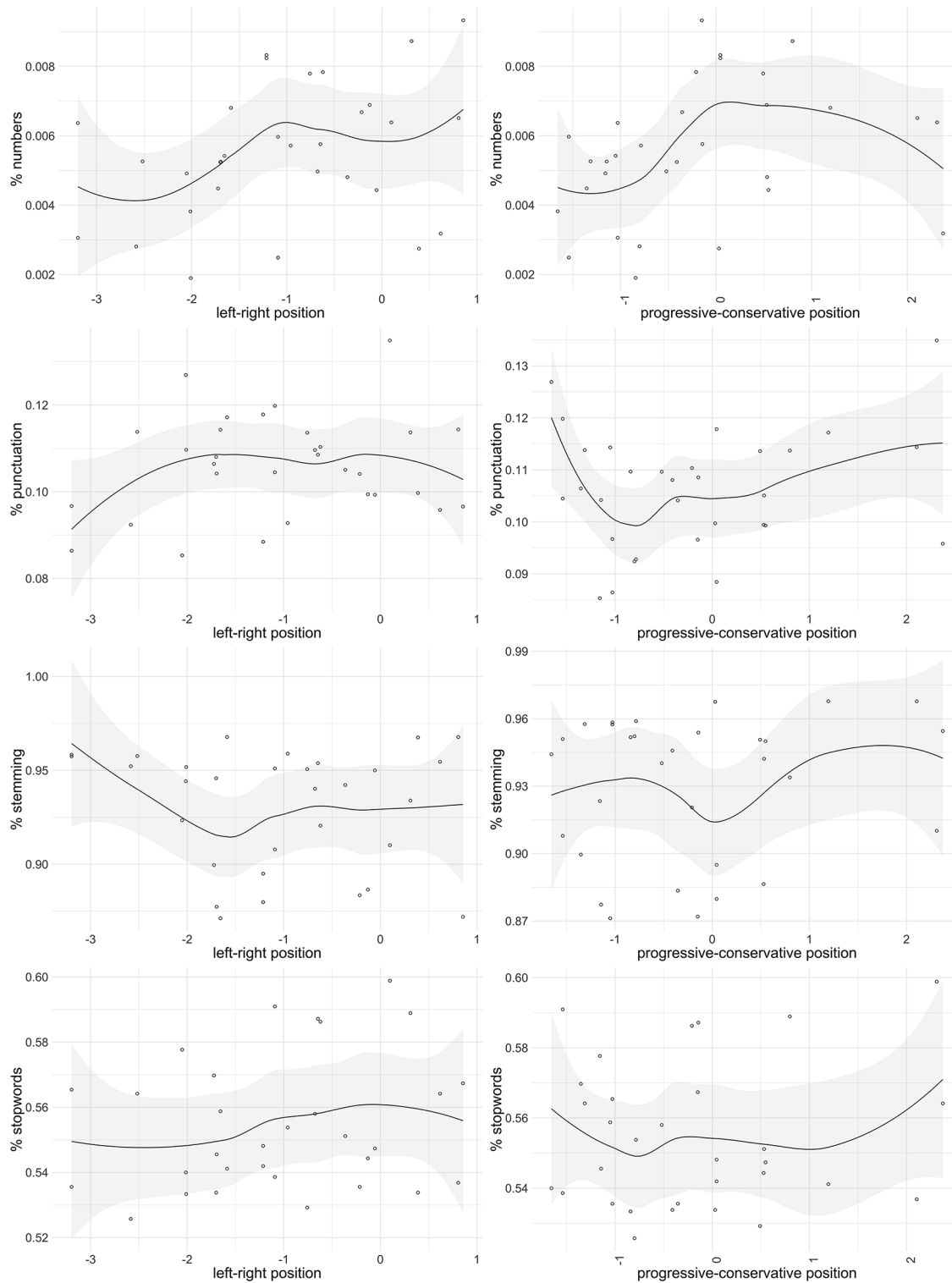


Figure 2. Average use of numbers, punctuation, stop words by politicians in the European Union, sorted by ideological left-right and progressive-conservative positions.

Note. This figure displays the average scores on each four preprocessing dimensions (numbers, punctuation, stemming and stop words) for EU politicians (heads of government and group leaders in the European Parliament), sorted by left-right and progressive-conservative ideology.

These results serve as evidence that these preprocessing steps are indeed not “ideologically neutral”, but the question remains what implications this has for unsupervised models like *Wordfish* in a large corpus like EUSpeech. In order to explore this question, we fitted 5 times a *Wordfish* model on our corpus: once with stop words removed; once with stemming; once with numbers removed; once with interpunction removed; and once without any of these preprocessing steps included. We excluded words that appeared among fewer than 10 speakers. The results are in Figure 3, which displays perfectly positive Spearman’s rank correlations between *Wordfish* positions estimated without any preprocessing steps and *Wordfish* positions estimated with each one of the four preprocessing steps respectively ($p = 0.98$ for removing stopwords and $p = 0.99$ for the other 3 preprocessing steps). The reason for this is likely the change in the number of features following each preprocessing step is modest compared to the size of the total corpus: for example, removing stop words removes only 139 features, a number that is swamped by the total number of features (5,564 and 5,425 respectively).^{xi} The same goes for the other preprocessing steps: their impact on the corpus on which the *Wordfish* estimates are based is negligible when compared against the total number of features: when removing numbers, the total number of features is 5,463; when removing punctuation, the total number of features is 5,545; when stemming the corpus, the total number of features decreases considerably more to 3,840 but this has no impact on *Wordfish* positions either.

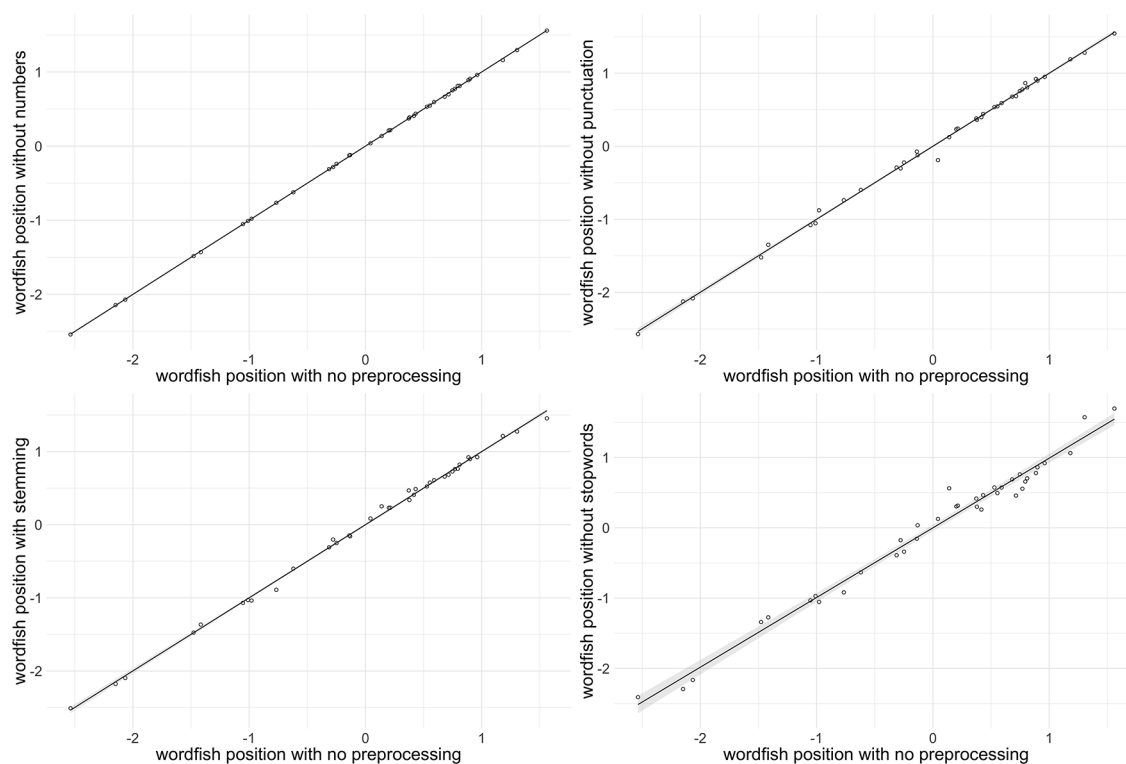


Figure 3. *Wordfish* estimates based on speeches by politicians in the European Union, with and without four preprocessing steps applied to the corpus.

Note. This figure displays *Wordfish* estimates based on speeches by politicians in the European Union, with and without four common preprocessing steps applied to the corpus. All four preprocessing steps (removal of numbers, punctuation, stop words, and stemming) appear not to have an influence on the estimated *Wordfish* positions.

From this we conclude that the amount of data (3,301 speeches from 31 speakers) overrides the potentially detrimental impact of preprocessing on subsequent scaling results.

Political scientists have recently been alerted to the dangers of preprocessing for unsupervised models (Denny & Spirling, 2018). We showed some evidence that seemingly random preprocessing steps (such as taking out stop words, numbers and punctuation, as well as stemming) correlate with stable characteristics like left-right ideology and liberal-conservative ideology (see Figure 2). Denny and Spirling (2018) demonstrate that these preprocessing steps also produce substantively different results. Using a much larger corpus we do not find that preprocessing steps influence estimated *Wordfish* positions (see Figure 3). The likely reason for this is that these preprocessing steps have a very small impact on the total number of features in this large corpus. In terms of concrete advice, we propose that when applying preprocessing steps on a small corpus, researchers are well-advised to consider the lessons from Denny and Spirling (2018) by using *preText* and average results across different model specifications. We also note that in a larger corpus, preprocessing steps may be less influential. We think that moving forward, work in political psychology on stable language patterns can inform a theory of *when* and *why* preprocessing matters exactly.

Analyzing Text

Can multifaceted concepts such as policy positions, topics, sentiment, complexity and personality be extracted from text? We think they can. Yet the problem is that typically we extract all these concepts simultaneously, while intending to extract just one. In other words, catching one construct may come with *bycatch* of another construct. This has implications for the substantive interpretation for the construct under study. This section illustrates this issue further, through a conceptual and an empirical example.

Let us start with a conceptual illustration of construct by-catch, using topic models and scaling models as an example. Scaling models use word co-occurrences between texts to place them on a single policy dimension. The more words co-occur between two texts, the closer they are placed on this dimension. This approach assumes that this dimension is what drives the dissimilarities between texts (Grimmer & Stewart, 2013), which requires that politicians talk differently about similar topics. For example, the sentence “we will raise unemployment benefits”, is very similar to “we will not raise unemployment benefits” and it is more dissimilar to an ideologically similar sentence like “levels of unemployment assistance should be increased”. For scaling to work the assumption is that a right-wing politician does not say “we will not raise unemployment benefits”, but instead says “handouts to the poor should be slashed”. Topic models, on the other hand, use word co-occurrences to classify text into one or more issues or topics. This builds on the assumption that different actors use identical words when talking about a topic. Taking the example of unemployment benefits, a topic model would place these sentences in separate topics, one characterized by words like “unemployment” and “benefits”, and the other by “handouts” and “poor”. The drivers of politicians’ language use are thus of crucial importance. If politicians only emphasize issues on which they are perceived as strong (Budge & Farlie, 1983), then political texts vary in their language as a result of different parties talking about different topics. However, when parties engage with each other on issues (e.g., Green-Pedersen & Mortensen, 2015) then scaling only works if these parties use language on similar topics that is different enough. In practice, parties sometimes engage with and sometimes avoid issues (Green-Pedersen & Mortensen, 2015). If the scaling procedure finds dissimilarities between parties it is because they talk about different issues or because they take different positions on the same issues. Only in the latter scenario do scaling methods help distinguish policy positions. In the former scenario scaling methods would distinguish between different topics instead. Thus, meaningful interpretation of scaling models depends on potential by-catch of topics.

As a further illustration of the issue of construct by-catch, we provide an example involving scaling and sentiment. We take speeches longer than 200 words (again from the EUSpeech data) that were originally delivered in English by prime ministers from nine EU member states and party leaders in the European Parliament ($n = 3,301$) which we aggregate across 31 speakers. Using Quanteda (Benoit & Nulty, 2016) we fit on these speeches a *Wordfish* scaling model and collect estimated speech positions. We also collect for each speech the percentage of negative and positive words using the Lexicoder Sentiment Dictionary (Young & Soroka, 2012) by taking the number of sentiment word occurrences and dividing by the total number of words in that speech. We then aggregate both position and emotion words at the speaker level.

If scaling is about position alone then there should be no relationship between *Wordfish* positions and use of sentiment words. Yet this is not what we observe (see Figure 4). Instead, speakers on one end of the underlying dimension use more positive words than speakers on the other end ($r = -0.34$). For example, about 10 percent of words used by English Prime Minister David Cameron are positive, whereas on the other side of the *Wordfish* dimension, about 8 percent of words of Greek Prime Minister Papademos are positive.^{xii} The relationship between the use of negative emotions with *Wordfish* positions is less pronounced ($r = 0.14$).^{xiii}

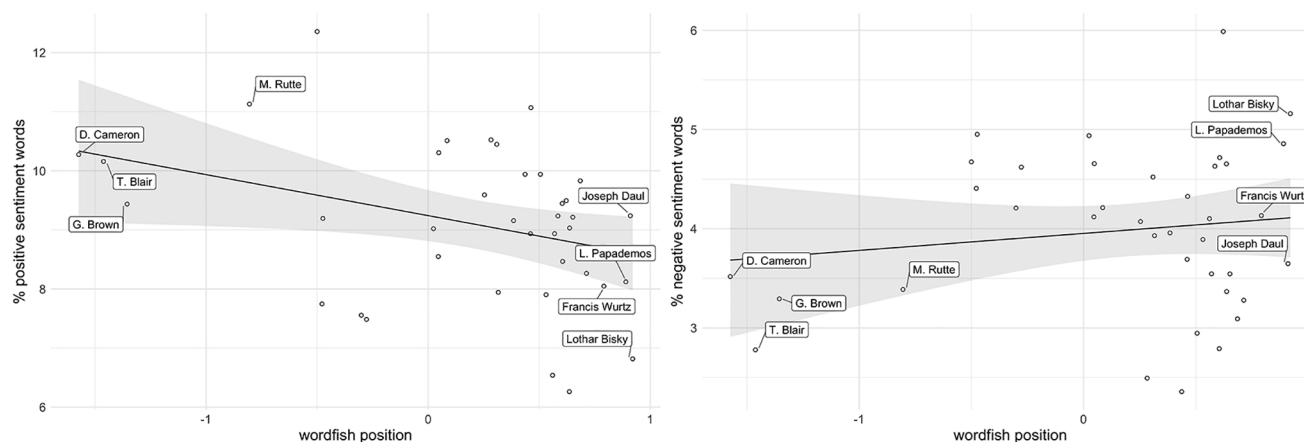


Figure 4. Use of sentiment and estimated *Wordfish* positions of speeches by politicians in the European Union.

Note. These scatterplots denote the average use of negative sentiment (left) and average use of positive sentiment (right) over the range of the estimated average *Wordfish* position of heads of government and MEP group leaders. It shows that *Wordfish* scores and the use of positive sentiment are correlated with each other.

The examples in this section show that words on which co-occurrence models like *Wordfish* are based may not have anything to do with policy positions, other than that they are correlated with them. This makes it difficult to substantively interpret positions on the dimension that *Wordfish* estimates. For example, the conclusion that “speaker A is more left-wing than speaker B” may be based on the fact that speaker A uses more sentiment words than speaker B. This may reflect real policy differences - speaker A is more sentimental about the topic - but it may also reflect real personality differences - speaker A is more emotional than speaker B. In any case, sentiment and position become blurred and we do not know which conclusion is justified. The underlying issue is one of construct by-catch. To distinguish between different constructs the analyst could apply a predictive validity criterion: if estimated policy positions are known to correlate with the use of emotion words, the analyst will need to account for these emotion words in subsequent analyses. If this alters the results, this should also alter the substantive conclusions. We thus encourage researchers – both in political psychology and political science – to be aware of

construct *by-catch* and check their results against other possible explanations. A popular way to cross-validate results is to use coders and train data. This, however, is not a feasible option for many researchers, especially for those working outside the English language context. Instead, we propose that researchers cross-validate their findings using other tools from the automated text analysis toolkit. In our example, we combined scaling and sentiment analyses. Other options are to use different dictionary methods or topic modeling.

Conclusion

Text is not a silver bullet for learning about politics and psychology. We have highlighted a number of issues to consider for each text as data project around four steps in the research process: (i) sampling text; (ii) authorship as meta data; (iii) preprocessing text; (iv) analyzing text.

Our discussion of these issues has reflected our optimism about the possibilities of text as a data source in both political psychology and political science, and we highlighted a set of guidelines which we summarize in [Table 1](#).

Table 1

Guidelines for Text as Data Projects in Political Psychology and Political Science

Guidelines
1. When collecting a corpus, use similar text sources to the extent possible. When multiple text sources are the only option, account for this in the analysis.
2. Get to know your data. How did a text come about? Who was involved? Incorporate this information in the analysis.
3. Consider in what ways preprocessing steps can correlate with stable speaker characteristics. Average results across preprocessing steps, particularly when working with a small corpus.
4. Use multiple methods to evaluate the possibility of construct by-catch when analyzing text.

We would like to conclude with a few observations. First, our disciplines (political psychology and political science) have come to expect much from the quality of, say, surveys or experimental data and it would be good that we apply that same rigorous standard to text. We encourage analysts in psychology and political science to be mindful of the quality of the texts they use, with an eye towards the construct under study, and we made some suggestions to that end. Second, our ability to extract those constructs will require us to think about data theory ([Jacoby, 1991](#)), research design and preprocessing steps. For example, preprocessing steps may have substantive implications when text is mined for constructs like personality (which research political psychology has shown to be reflected in stable language patterns). Third, we want to emphasize the importance of further theory-building and concept development. As it currently stands, the literature converges on the use of existing “gold standards” like the Linguistic Inquiry and Word Count (LIWC) dictionary ([Pennebaker et al., 2015](#)) or the ANEW (Affective Norms for English Words). Although these measurement instruments are highly valuable, analysts should keep questioning them – and build alternatives – to avoid running the risk of depending too much on them.

In their early review paper, [Grimmer and Stewart \(2013\)](#) urged researchers using text as data to “validate, validate, validate” the outputs of their models. We believe that an important way of doing so for researchers working on text as data projects is by integrating different perspectives in their work. For example, political scientists could

learn from political psychologists about how individual characteristics are reflected in stable language patterns among politicians, whereas political scientists could learn from political psychologists how the political context (e.g., the dynamics a political campaign or the intended audience of a speech) pressures these politicians into changing their language use. The promise that text as data holds for political psychology and political science will be bolstered with more cross-fertilization – theoretically and empirically – between both disciplines.

Notes

i) In [Appendix A](#) we provide for interested readers a brief overview of popular applications of automated text analysis methods: *topical content, sentiment, scaling, integrative and cognitive complexity, and personality*.

ii) We thank Richard Slatcher for making their data available.

iii) The data and scripts to replicate the results in this paper can be found on Harvard's Dataverse: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/2PNZNU>

iv) Authorship assessment made the news recently with publication of the 'inside the White House anonymous resistance' op-ed directed against Trump: <https://www.nytimes.com/2018/09/05/opinion/trump-white-house-anonymous-resistance.html>

v) Their R package *preText* implements this procedure.

vi) Linguists distinguish between content words and function words. Content words such as nouns, verbs, adverbs and adjectives have meaning. Function words such as prepositions, pronouns and conjunctions explain or create grammatical or structural relationships into which the content words may fit.

vii) For example, [Denny and Spirling \(2018\)](#) analysis of variation in *Wordfish* positions and preprocessing steps consists of just 8 election manifestos.

viii) This is relatively short. For example, the TM package in R ([Feinerer & Kurt, 2017](#)) contains a stop word list of 488 words.

ix) This includes items about EU, immigration, traditional values and morality, as well as items on international solidarity.

x) This includes items in the economic topic category, except for the "economic goals" and "technology and infrastructure: positive" items, as well as items about welfare and education.

xi) Based on [Figure 3](#), the strongest predictor of *Wordfish* positions appears to be mother tongue with native English speakers located on one side of the spectrum and native Spanish speakers located on the other side.

xii) This is still considerably more positive than Gabriele Zimmer, head of the Nordic Green Left in the European Parliament. In her speeches only 6.4 percent of words are positive.

xiii) It is interesting to note that even though we only included those speeches that were delivered in English – in addition to sentiment – *Wordfish* seems to distinguish between different nationalities as well: among one end of the spectrum we find English speakers like Gordon Brown, Cameron and Blair, whereas on the side we find, among others, French European Parliament Group leaders Francis Wurtz and Joseph Daul.

Funding

For conducting this research, Schoonvelde and Schumacher received funding from the European Union's Horizon2020 research and innovation program under grant agreement No 649281, EUENGAGE.

Competing Interests

The authors have declared that no competing interests exist.

Acknowledgments

The authors would like to thank Frederik Hjorth, Fabian Neuner, Roland Kappe, and Mariken van der Velden, as well as attendants of the Dutch Political Psychology Meeting in The Hague and two anonymous reviewers for helpful comments on earlier drafts of this paper.

References

- Airoldi, E. M., Fienberg, S. E., & Skinner, K. K. (2007). Whose ideas? Whose words? Authorship of Ronald Reagan's radio addresses. *PS: Political Science & Politics*, *40*, 501-506. <https://doi.org/10.1017/S1049096507070874>
- Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)* (pp. 579–586). Vancouver, Canada: Association for Computational Linguistics.
- Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. W. (2005). *Lexical predictors of personality type*. Paper presented at the Joint Annual Meeting of the Interface and The Classification Society of North America, St. Louis, MI, USA. Retrieved from <https://pdfs.semanticscholar.org/e70c/7a9a8568203921a9a582626cbaf97cc6c2d8.pdf>
- Bakker, B. N. (2017). Personality traits, income, and economic ideology. *Political Psychology*, *38*, 1025-1041. <https://doi.org/10.1111/pops.12349>
- Benoit, K., & Laver, M. (2007). Benchmarks for text analysis: A response to Budge and Pennings. *Electoral Studies*, *26*, 130-135. <https://doi.org/10.1016/j.electstud.2006.04.001>
- Benoit, K., & Nulty, P. (2016). Quanteda: Quantitative analysis of textual data (version 0.9) [R package]. Retrieved from <https://quanteda.io>
- Bischof, D., & Senninger, R. (2018). Simple politics for the people? Complexity in campaign messages and political knowledge. *European Journal of Political Research*, *57*, 473-495. <https://doi.org/10.1111/1475-6765.12235>
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*, 77-84. <https://doi.org/10.1145/2133806.2133826>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.
- Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*, 8-23. <https://doi.org/10.1080/21670811.2015.1096598>
- Bradley, M. M., & Lang, P. J. (1999). *Affective Norms for English Words (ANEW): Stimuli, instruction manual and affective ratings* (Tech. Report C-1). Gainesville, FL, USA: University of Florida, Center for Research in Psychophysiology.
- Budge, I., & Farlie, D. (1983). *Explaining and predicting elections: Issue effects and party strategies in twenty-three democracies*. London, United Kingdom: Allen & Unwin.
- Calvo, R. A., & Mac Kim, S. (2013). Emotions in text: Dimensional and categorical models. *Computational Intelligence*, *29*, 527-543. <https://doi.org/10.1111/j.1467-8640.2012.00456.x>
- Caprara, G., Francescato, D., Mebane, M., Sorace, R., & Vecchione, M. (2010). Personality foundations of ideological divide: A comparison of women members of parliament and women voters in Italy. *Political Psychology*, *31*, 739-762. <https://doi.org/10.1111/j.1467-9221.2010.00780.x>
- Collins, P. (2012). *The art of speeches and presentations: The secrets of making people remember what you say*. Chichester, United Kingdom: John Wiley and Sons.

- Conway, L. G., III, Conway, K. R., Gornick, L. J., & Houck, S. C. (2014). Automated integrative complexity. *Political Psychology*, 35, 603-624. <https://doi.org/10.1111/pops.12021>
- De Lange, S., & van Erkel, P. (2013). *Estimating the influence of coalition partners on coalition agreements using wordfish and wordscores*. Paper presented at the 7th ECPR General Conference, Bordeaux, France.
- Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26, 168-189. <https://doi.org/10.1017/pan.2017.44>
- Dietrich, B. J., Lasley, S., Mondak, J. J., Remmel, M. L., & Turner, J. (2012). Personality and legislative politics: The Big Five trait dimensions among US state legislators. *Political Psychology*, 33, 195-210. <https://doi.org/10.1111/j.1467-9221.2012.00870.x>
- Dille, B. (2000). The prepared and spontaneous remarks of Presidents Reagan and Bush: A validity comparison for at-a-distance measurements. *Political Psychology*, 21, 573-585. <https://doi.org/10.1111/0162-895X.00205>
- Druckman, J. N., Kuklinski, J. H., & Sigelman, L. (2009). The unmet potential of interdisciplinary research: Political psychological approaches to voting and public opinion. *Political Behavior*, 31, 485-510. <https://doi.org/10.1007/s11109-009-9092-2>
- Feinerer, I., & Kurt, H. (2017). Package 'tm' (version 0.7-5) [R package]. Retrieved from <https://cran.r-project.org/web/packages/tm/tm.pdf>
- Gemenis, K. (2013). What to do (and not to do) with the comparative manifestos project data. *Political Studies*, 61(S1), 3-23. <https://doi.org/10.1111/1467-9248.12015>
- Green-Pedersen, C., & Mortensen, P. B. (2015). Avoidance and engagement: Issue competition in multiparty systems. *Political Studies*, 63, 747-764. <https://doi.org/10.1111/1467-9248.12121>
- Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18, 1-35. <https://doi.org/10.1093/pan/mpp034>
- Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21, 267-297. <https://doi.org/10.1093/pan/mps028>
- Hermann, M. G. (1980). Explaining foreign policy behavior using the personal characteristics of political leaders. *International Studies Quarterly*, 24, 7-46. <https://doi.org/10.2307/2600126>
- Hermann, M. G. (2005). Assessing leadership style: A trait analysis. In J. M. Post (Ed.), *The psychological assessment of political leaders* (pp. 178-212). Ann Arbor, MI, USA: University of Michigan Press.
- Iacobelli, F., Gill, A. J., Nowson, S., & Oberlander, J. (2011). Large scale personality classification of bloggers. In S. D'Mello, A. Graesser, B. Schuller, & J. C. Martin (Eds.), *Affective computing and intelligent interaction* (pp. 568-577). Berlin, Germany: Springer.
- Jacoby, W. G. (1991). *Data theory and dimensional analysis*. Newbury Park, CA, USA: Sage.
- Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated readability index, Fog count and Flesch reading ease formula) for navy enlisted personnel* (Research Branch report 8-75). Institute for Simulation and Training, University of Central Florida, Florida, USA.
- Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97, 311-331. <https://doi.org/10.1017/S0003055403000698>

- Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology, 90*, 862-877. <https://doi.org/10.1037/0022-3514.90.5.862>
- Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence, 29*, 436-465. <https://doi.org/10.1111/j.1467-8640.2012.00460.x>
- Noonan, P. (1998). *On speaking well: How to give a speech with style, substance, and clarity*. New York, NY, USA: Regan Books.
- Oberlander, J., & Gill, A. J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes, 42*, 239-270. https://doi.org/10.1207/s15326950dp4203_1
- Pennebaker, J. W. (2011). *The secret life of pronouns: How our words reflect who we are*. New York, NY, USA: Bloomsbury.
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX, USA: University of Texas at Austin.
- Pennebaker, J. W., & King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*, 1296-1312. <https://doi.org/10.1037/0022-3514.77.6.1296>
- Proksch, S.-O., & Slapin, J. B. (2009). How to avoid pitfalls in statistical analysis of political texts: The case of Germany. *German Politics, 18*, 323-344. <https://doi.org/10.1080/09644000903055799>
- Ramey, A. J., Klingler, J. D., & Hollibaugh, G. E. (2019). Measuring elite personality using speech. *Political Science Research and Methods, 7*, 163-184. <https://doi.org/10.1017/psrm.2016.12>
- Renshon, J. (2009). When public statements reveal private beliefs: Assessing operational codes at a distance. *Political Psychology, 30*, 649-661. <https://doi.org/10.1111/j.1467-9221.2009.00718.x>
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., . . . Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*, 1064-1082. <https://doi.org/10.1111/ajps.12103>
- Schoonvelde, M., Brosius, A., Schumacher, G., & Bakker, B. N. (2019). Liberals lecture, conservatives communicate: Analyzing complexity and ideology in 381,609 political speeches. *PLoS ONE, 14*(2), Article e0208450. <https://doi.org/10.1371/journal.pone.0208450>
- Schumacher, G., Schoonvelde, M., Dahiya, T., & De Vries, E. (2016). *EUSpeech* (Harvard Dataverse, V1) [Dataset]. Retrieved from <https://dataverse.harvard.edu/dataverse/euspeech>
- Schumacher, G., Schoonvelde, M., Traber, D., Dahiya, T., & De Vries, E. (2016). EUSpeech: A new dataset of EU elite speeches. In *Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016)* (pp. 75–80). Dubrovnik, Croatia, 14–16 July 2016.
- Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science, 52*, 705-722. <https://doi.org/10.1111/j.1540-5907.2008.00338.x>
- Slatcher, R. B., Chung, C. K., Pennebaker, J. W., & Stone, L. D. (2007). Winning words: Individual differences in linguistic style among U.S. presidential and vice presidential candidates. *Journal of Research in Personality, 41*, 63-75. <https://doi.org/10.1016/j.jrp.2006.01.006>
- Strapparava, C., & Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing* (pp. 1556-1560). Fortaleza, Ceara, Brazil.

- Swaim, B. (2015). *The speechwriter: A brief education in politics*. New York, NY, USA: Simon & Schuster.
- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*, 24-54. <https://doi.org/10.1177/0261927X09351676>
- Tetlock, P. E. (1981). Pre- to post-election shifts in presidential rhetoric: Impression management or cognitive adjustment. *Journal of Personality and Social Psychology, 41*, 207-212. <https://doi.org/10.1037/0022-3514.41.2.207>
- Thoemmes, F. J., & Conway, L. G., III. (2007). Integrative complexity of 41 U.S. Presidents. *Political Psychology, 28*, 193-226. <https://doi.org/10.1111/j.1467-9221.2007.00562.x>
- Volkens, A., Lehmann, P., Matthieß, T., Merz, N., Regel, S., & Werner, A. (2016). *The manifesto data collection* (Manifesto project [MRG/CMP/MARPOR], Version 2015) [Dataset]. Berlin, Germany: Wissenschaftszentrum Berlin für Sozialforschung/WZB.
- Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science, 20*, 529-544. <https://doi.org/10.1146/annurev-polisci-052615-025542>
- Winter, D. G. (2005). Things I've learned about personality from studying political leaders at a distance. *Journal of Personality, 73*, 557-584. <https://doi.org/10.1111/j.1467-6494.2005.00321.x>
- Winter, D. G., Hermann, M. G., Weintraub, W., & Walker, S. G. (1991). The personalities of Bush and Gorbachev at a distance: Follow-up on predictions. *Political Psychology, 12*, 457-464. <https://doi.org/10.2307/3791754>
- Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication, 29*, 205-231. <https://doi.org/10.1080/10584609.2012.671234>

Appendices

Appendix A: Text as Data Applications

For readers who are new to these methods, this appendix briefly discusses popular applications of measuring constructs from text: *topics, ideological dimensions, sentiment, complexity, personality*.

Topic models use the co-occurrence of words across texts to identify topics. A text is assumed to cover a single topic or multiple topics (mixture models). A prominent example of topic models is Latent Dirichlet Allocation. LDA assumes that each text can be represented as a mixture of topics. Topics are considered to be a probability distribution over words (Blei, 2012; Blei, Ng, & Jordan, 2003; Boumans & Trilling, 2016). LDA is a generative model, which takes the words in each document as input, and from that input plus some statistical assumptions infers the latent topical structure—the topics, and their distribution per document and across the corpus. Since LDA is an unsupervised method, the researcher has to infer the meaning of each topic.

Scaling models place texts from say political parties on a single underlying dimension. This is typically a left-right dimension. *Wordscores* (Laver, Benoit, & Garry, 2003) and *Wordfish* (Proksch & Slapin, 2009; Slapin & Proksch, 2008) are two examples of ideological scaling models. *Wordscores* is a supervised approach to scaling political text. The researcher identifies political texts with known ideological characteristics, the “reference” texts. The distribution of words used in new documents is then compared to the distribution of words in these reference texts (Laver et al., 2003). The output is a similarity score, comparing how much alike the document is to the reference texts. This approach assumes clear a priori knowledge of the underlying dimension in the text. *Wordfish*, however, extracts substantively relevant quantities in an unsupervised manner (Slapin & Proksch, 2008). In the *Wordfish* approach documents are given positions on a latent dimension. Unlike *Wordscores*, the researcher

cannot be a priori certain of the underlying dimension based on a set of reference texts, but the research can select anchors (words or documents) to retrieve the meaning of the latent dimension that *Wordfish* produces.

Sentiment analysis is applied in various disciplines to extract the level of sentiment from text. Sentiment analysis can assess the level and direction of emotional arousal in a text, as well as the presence of specific discrete emotions (usually: *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*). Much of this work is based on dictionaries, such as the Affective Norms for English Words (ANEW) (Bradley & Lang 1999), Linguistic Inquiry of Word Count (LIWC) (Pennebaker, Boyd, Jordan, & Blackburn, 2015), Lexicoder (Young & Soroka, 2012) and the NRC Word-Emotion Association Lexicon (Mohammad & Turney, 2013).^{XIV} A simple application of ANEW is to score text on each of these dimensions, through a tally of words in the text that also appear in the ANEW word list. The NRC Word-Emotion Association Lexicon is a crowd-sourced lexicon which distinguishes between *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, *trust* (Mohammad & Turney, 2013). More advanced computational applications in computer science (e.g., Calvo & Mac Kim, 2013; Strapparava & Mihalcea, 2008) aim to categorize text — for example, news headlines — with regards to their categorical emotional content.

Complexity is another feature of text that is occasionally analyzed in political science (Bischof & Senninger, 2018; Conway, Conway, Gornick, & Houck, 2014; Hermann, 2005; Schoonvelde, Brosius, Schumacher, & Bakker, 2019; Tetlock 1981; Thoemmes & Conway, 2007). In its most basic form, complexity is measured as the number of words per sentence, and the length of the words in the sentences. The Flesch-Kincaid readability score is an example of such a complexity measure (Kincaid, Fishburne, Rogers, & Chissom, 1975). This measure was initially used by education researchers to gauge the readability of a text. But there are several more measures of complexity such as integrative complexity and cognitive complexity. Cognitive complexity measures the degree of multidimensional, differentiated thinking revealed in a text. If a speaker or author gives several perspectives on a given topic, a text becomes conceptually more complex (Hermann, 2005; Pennebaker & King, 1999). Exclusion words such as ‘but’, ‘without’ and ‘exclude’ signify differentiation between perspectives, as well as conjunctions such as ‘also’, ‘and’ and ‘although’ (Tausczik & Pennebaker, 2010). Similarly, Hermann (1980) looks for words such as ‘may’, ‘possibly’, ‘sometimes’ (high complexity), and ‘always’, ‘only’ and ‘without a doubt’ (low complexity). Other work further distinguishes the level of integrative complexity by distinguishing between elaborative and dialectical forms of complexity (Thoemmes & Conway, 2007).

A recent application even uses texts to identify the *personality* of a politician (Ramey, Klingler, & Hollibaugh, 2019). There is an older more qualitative tradition that also identified personality from text (Hermann, 2005; Winter, 2005). Initially, psychologists studying personality and language patterns found recurring patterns of linguistic habits and specific personality traits (see e.g., Pennebaker & King, 1999). These recurring patterns have helped computer scientists to build models that can predict personality characteristics based on text. By comparing written texts to personality surveys, computer scientists have been able to train models able to predict personality of bloggers (Iacobelli, Gill, Nowson, & Oberlander, 2011), US congressmen (Ramey et al., 2019) and students (Argamon, Dhawle, Koppel, & Pennebaker, 2005).

Table A1

Key Citations of Popular Applications of Automated Text Analysis Methods

Construct	Key Citations
Complexity	Hermann (1980, 2005); Tetlock (1981)
Emotions	Alm et al. (2005); Calvo & Mac Kim (2013); Mohammad & Turney (2013)
Personality	Mehl et al. (2006); Oberlander & Gill (2006); Pennebaker & King (1999)
Scaling	Benoit & Laver (2007); Proksch & Slapin (2009); Slapin & Proksch (2008)
Topics	Blei (2012); Blei et al. (2003); Roberts et al. (2014)

Appendix B: OLS Regression of Preprocessing Steps on Ideology

Table B1

OLS Model of Level of Stop Word Use on Speaker Ideology

Variable	(Model 1)	(Model 2)	(Model 3)
Left Right Position	-0.002*** (0.001)		-0.002* (0.001)
Progressive Conservative Position		-0.002* (0.001)	0.001 (0.001)
Left Right X Progressive Conservative			0.002** (0.001)
CONSTANT	0.529*** (0.003)	0.532*** (0.003)	0.527*** (0.003)
Observations	3,157	3,157	3,157
R-squared	0.137	0.135	0.139
Adjusted R-squared	0.133	0.131	0.135
Residual standard error	0.030 (df = 3144)	0.030 (df = 3144)	0.030 (df = 3142)
F statistic	41.424*** (df = 12; 3144)	40.766*** (df = 12; 3144)	36.166*** (df = 14; 3142)

Note. Standardized coefficients, standard errors in parentheses.

Table B2

OLS Model of Proportion of Numbers on Speaker Ideology

Variable	(Model 1)	(Model 2)	(Model 3)
Left Right Position	-0.0001 (0.0001)		-0.0003+ (0.0002)
Progressive Conservative Position		0.0001 (0.0001)	0.0001 (0.0002)
Left Right X Progressive Conservative			-0.0004*** (0.0001)
CONSTANT	0.002** (0.001)	0.002*** (0.0005)	0.002*** (0.001)
Observations	3,157	3,157	3,157
R-squared	0.063	0.063	0.069
Adjusted R-squared	0.060	0.060	0.065
Residual standard error	0.006 (df = 3144)	0.006 (df = 3144)	0.005 (df = 3142)
F statistic	17.724*** (df = 12; 3144)	17.708*** (df = 12; 3144)	16.624*** (df = 14; 3142)

Note. Standardized coefficients, standard errors in parentheses.

Table B3

OLS Model of Proportion of Unique Words on Speaker Ideology

Variable	(Model 1)	(Model 2)	(Model 3)
Left Right Position	0.008*** (0.001)		0.003* (0.001)
Progressive Conservative Position		0.013*** (0.001)	0.014*** (0.001)
Left Right X Progressive Conservative			0.006*** (0.001)
CONSTANT	0.968*** (0.004)	0.962*** (0.004)	0.958*** (0.004)
Observations	3,157	3,157	3,157
R-squared	0.346	0.358	0.374
Adjusted R-squared	0.343	0.356	0.372
Residual standard error	0.041 (df = 3144)	0.041 (df = 3144)	0.040 (df = 3142)
F statistic	138.489*** (df = 12; 3144)	146.094*** (df = 12; 3144)	134.266*** (df = 14; 3142)

Note. Standardized coefficients, standard errors in parentheses.

Table B4

OLS Model of Level of Punctuation on Speaker Ideology

Variable	(Model 1)	(Model 2)	(Model 3)
Left Right Position	0.003*** (0.0004)		0.002** (0.0005)
Progressive Conservative Position		0.004*** (0.0004)	0.003*** (0.001)
Left Right X Progressive Conservative			0.0004 (0.0003)
CONSTANT	0.116*** (0.002)	0.113*** (0.001)	0.115*** (0.002)
Observations	3,157	3,157	3,157
R-squared	0.148	0.151	0.154
Adjusted R-squared	0.145	0.148	0.151
Residual standard error	0.016 (<i>df</i> = 3144)	0.016 (<i>df</i> = 3144)	0.016 (<i>df</i> = 3142)
F statistic	45.594*** (<i>df</i> = 12; 3144)	46.735*** (<i>df</i> = 12; 3144)	40.996*** (<i>df</i> = 14; 3142)

Note. Standardized coefficients, standard errors in parentheses.