The Nuts and Bolts of Automated Text Analysis. Comparing Different Document Pre-Processing Techniques in Four Countries

Greene, Z.; Ceron, A.; Schumacher, G.; Fazekas, Z.

Publication date
2016

Document Version
Final published version

The Nuts and Bolts of Automated Text Analysis:

Comparing Different Document Pre-Processing Techniques in Four Countries

Zac Greene
University of Strathclyde
Andrea Ceron
University of Milan
Gijs Schumacher
University of Amsterdam
& Zoltan Fazekas
University of Oslo

Abstract

Automated text analytic techniques have taken on an increasingly important role in the study of parties and political speech. Researchers have studied manifestos, speeches in parliament, and debates at party national meetings. These methods have demonstrated substantial promise for measuring latent characteristics of texts. In application, however, scaling models require a large number of decisions on the part of the researcher that likely hold substantive implications for the analysis. Past researchers proposed discussion of these implications, but there is no clear prescription or systematic examination of these choices with the goal of establishing a set of best practices based on their implications for speeches at parties' national meetings in a comparative setting. We examine the implications of these choices with data from intra-party meetings in Germany, Italy, the Netherlands, and prime minister speeches in Denmark. We conclude with considerations for those undertaking political text analyses.

Automated text analysis methods offer substantial opportunities to develop and test political science theories. These tools have been used to explain diverse topics such as MPs' behavior in parliament (Schonhardt-Bailey 2006; Klemmensen et al. 2007; Quinn et al. 2010; Proksch and Slapin 2012; Eggers and Spirling 2014), policy positions (Laver et al. 2003; Slapin and Proksch 2008; Proksch and Slapin 2009), legislator press releases (Grimmer 2010), central bank statements (Baerg 2014; Baerg and Lowe 2015), news reports (Van Atteveldt et al. 2008; Coscia and Rios 2012; Stockmann 2012), intra-party divisions (Ceron 2012, 2013, 2014; Greene and Haber 2014; Lo et al. 2014), social media (King et al. 2013; Barbera 2015) and US government treaties with Native American tribes (Spirling 2012). Alternate approaches to scale the latent preferences of actors' have been particularly useful for some applications in comparative politics research (e.g. Laver et al. 2003; Slapin and Proksch 2008). This path-breaking research comes with clear prescriptions for the efficient and valid use of automated text techniques in various settings (e.g. Laver et al. 2003; Lowe 2008; Monroe et al. 2008; Proksch and Slapin 2009; Grimmer and Stewart 2013; Lucas et al. 2015). In using these methods, analysts face substantial choices *prior* to the implementation of the primary analysis. For example, researchers often reduce linguistic complexity by removing uninformative stopwords and by stemming documents, and facilitate model estimation by removing rare (or very common) terms. Although these practices are common and uncontroversial in computer science and linguistics (e.g. Hollink et al. 2004; Manning et al. 2008), the potential substantive implications of these choices for frequently used scaling models applied to political texts are less clear (see Denny and Spirling 2016 for a similar approach). Differences between languages further complicate the formulation of best practices, especially for those engaged in comparative, cross national research (but see Lucas et al. 2015).

In this paper, we examine the consequences of these choices for frequently used models that estimate the latent positions of political actors from spoken and written texts (e.g. *Wordfish*). In particular, we review these practices and consider their implications for the estimated position of a document and the estimated uncertainty of that position. We propose that decisions related to the processing of political documents influence the consistency and the reliability of the results obtained from automated text analysis. Practices designed to

reduce linguistic complexity reduce the uncertainty associated with individual speeches by isolating the most informative words. We demonstrate these characteristics using party leader speeches and written texts of internal party debates in a variety of languages from Germany, Italy, and the Netherlands, and annual prime minister speeches from Denmark. We conclude that researchers should pay close attention to the processing decisions they make and encourage scholars to be transparent with the substantive implications of these decisions for their analyses.

Political Text Analysis and Dimensional Scaling

Text analysis has played a pioneering role in comparative research. Initial studies investigated political preferences and attention by means of large-scale human based coding projects. Through projects such as the Manifesto Research Group (e.g. Budge and Farlie 1983; Budge et al. 2001; Klingemann et al. 2005) and the Policy Agendas Project (Jones et al. 2009), individual coders used predefined topic labels to hand code documents like election manifestos and laws. These time and resource intensive projects have led to a large number of advancements in political research.

By the early 2000s, political science researchers had begun to explore automated techniques to exploit the burgeoning amount of textual information. In response to theoretical approaches emphasizing the effect of formal rules and actors' preferences, scholars proposed latent dimensional scaling models to extract speakers' relative preferences from political text. These models were often presented as an alternate data source to study the behavior of representatives behaviors in parliament where data is often lacking, incomplete or endogenous to the processes under evaluation (Laver et al. 2003; Loewenberg 2008; Proksch and Slapin 2008). Furthermore, they were also employed in order to overcome potential limitations and reliability issues related to human coding (Mikhaylov et al. 2011).

Following these advances, two approaches have shown particular promise for studying the preferences of parties and candidates: *Wordscores* and *Wordfish*. Both sets of scaling models have been used to explore the preferences of political actors and much debate

surrounds their use for scaling political documents and speeches (e.g. Benoit and Laver 2007 and 2008; Lowe 2008; Martin and Vanberg 2008; Proksch and Slapin 2009).

Proposed initially as a method to study the relative location of parliamentary leaders, *Wordscores* has since been applied to a broad range of political text such as election manifestos and parliamentary speech (e.g. Laver and Benoit 2002; Laver et al. 2003; Laver 2003; Giannetti and Laver 2005; Benoit and Herzog 2015). At its heart, *Wordscores* presents an *a priori* approach to studying political texts, falling into the category of supervised learning algorithms. The researcher identifies a set of political texts with known ideological characteristics. The distribution of words used in new documents is then compared to the distribution of words in an original set of "reference" texts (Laver et al. 2003). Each document is converted into a word frequency matrix that can be used to estimate the similarity of each document to the reference texts chosen.

While the reference texts used in the *Wordscores* approach provide clear *a priori* knowledge of the underlying dimension in the text, *Wordfish* measures the underlying differences more generally, extracting substantively relevant quantities in an unsupervised manner. In the *Wordfish* approach, documents are given positions on a latent dimension in which the dimension's direction can be determined by anchor documents. Unlike *Wordscores*, the researcher cannot be *a priori* certain of the underlying dimension based on a set of reference texts, but purposefully selected anchors inform the model so that the researchers know directionality of underlying conflict.

Although these two approaches have their differences, both rely on the (quite strong) assumption that the content of the political texts is predominantly ideological, and therefore informative of the policy position expressed by each actor (Grimmer and Stewart 2013). They also share another core element, namely their input, which is even more crucial in light of such 'ideological dominance' assumption. Which terms enter the word frequency matrices (that serve as the basis of the analyses) has fundamental implications for the results retrieved. Indeed, it has been argued that preprocessing, by removing words that are mainly related to stylistic features of the texts can allow to isolate more ideological words and mitigate some drawbacks of scaling approaches (Beauchamp 2011; Grimmer and Stewart 2013). In particular,

Beauchamp (2011) shows that removing technical language, which coincides more with party power than with ideology, significantly improves the estimates.

Few attempts have been made to explore the effect of pre-processing decisions on the outputs from each model. So far, this important feature of text analysis has rarely been addressed. Proksch and Slapin's (2008) discussion of stemming and stopwords reflects the most detailed discussion of their impact for the analysis and results, but they limit their focus to German election manifestos. Ultimately, they find evidence that both stemming and removing stopwords improve the identification of a document's position. As long as there are good reasons to believe that these implications may be very different for alternate forms of text (such as speeches) and different languages (Hollink et al. 2004), in the course of the paper we will adopt a comparative framework to highlight how pre-processing can affect text analysis.

Pre-Processing of Textual Data in Political Science

Scaling political texts to obtain estimates of policy position requires that researchers prepare texts for the analysis and, to do that in a proper way, researches must be familiar with the tools for carrying out this preparation (Benoit and Herzog 2015). Pre-processing is an important part of automated text analysis. This refers to the stage of analysis in which textual data are cleaned and prepared for quantitative analysis. While computer scientists historically dedicated substantial attention to these questions political scientists largely relegate these decisions to footnotes. Proksch and Slapin (2009) aptly suggest that researchers should be transparent about these choices, as pre-processing can have substantial implications for the success of automated text analysis. Hence, assessing whether particular pre-processing decisions carry systematic implications for the results and whether these implications are present uniformly across types of documents and languages, contributes to applied comparative politics using political texts.

Dimensional scaling in political science has largely taken two alternate approaches and these approaches came with different practices in terms of pre-processing. In the *Wordscores* tradition, researchers have largely chosen to avoid removing or altering the substance of documents. Laver et al. (2003) explain that they use all types of words because they seek to

"analyze texts in languages we do not understand" and these words "convey no useful information, but they do not systematically bias our results" (Laver et al. 2003, 315-316).

Scholars in the *Wordfish* approach often undertake some form of pre-processing. Proksch and Slapin (2009) find that their estimates are the most efficient when stopwords are excluded and after using a stemmer. Their results are consistent with research on document retrieval showing that stemmers are often useful for German language texts (Hollink et al. 2004). Although *Wordfish* can be run on the entire word count data matrix, Proksch and Slapin (2009b) additionally recommend to use only a subsample of words (unless the language has remained constant over time) and to remove words that are used very infrequently. Along facilitating across-time comparisons, this latter step facilitates model estimation by reducing the number of cells with zero word frequencies in the document-term matrix.

Following this research, we consider the implications of pre-processing. We focus on the consequences (in terms of *Wordfish* estimates) of the decision to remove stopwords, adopt word stemming, and reduce the sparsity of the word matrix. These practices are common points of discussion in computer science and linguistics, some of the steps applied, but less well known in political science (see for example, Manning 2008).

A common practice in machine learning and automated text analysis is to exclude certain words from the analysis. So called "stopwords" such as *why, to*, *it*, offer no additional meaning to an analysis focused on distinguishing a latent dimension of political conflict. Not all words are useful for extracting positions through scaling of political texts. By including them in the analysis, they reduce the efficiency of estimates and might potentially introduce bias based on certain rhetorical approaches (Manning et al. 2008). Scholars, however, are also reminded that stopword removal should be used with caution: as long as there is no 'one-size fits-all' list of words to be removed, researchers should inspect which words are worth removing and which words are not (Benoit and Herzog 2015).

Computer scientists often consider words the natural unit for analysis. While it makes sense for hand-coded texts to use full sentences or quasi-sentences (Budge et al. 2001), automated text analysis often focuses on the level of the word.[1] The issue with this approach is

[1] Other approaches such as k-grams use combinations of adjacent words to scale documents (Manning et al. 2008).

that words with similar meanings, but different endings often convey similar information about a text's latent position. The terms *democracy*, *democratic*, and *democratizing* likely all indicate a disposition towards the same substantive goal. Treating these words as distinct might over-estimate the similarity of two texts while increasing the estimates of uncertainty for both word and document weights.

Instead, stemming algorithms seek to better connect substantively similar words. In this process, stems are the sliced up pieces of a text (often a word), and these become the unit of analysis (Manning et al. 2008). The process of tokenization breaks sentences down into smaller pieces to extract meaningful differences between texts. The goal of stemming and lemmatization (a similar process based on a dictionary approach) is to reduce complex forms of words to their simplest root. On the one hand, stemming is a relatively simplistic approach that removes the end of the words. For example, stemming would cause the analysis to treat related words such as *unemployment* and *unemployed* equally by treating them as their word stem, *unemploy*. This results in converting a number of related words to a similar root. Lemmatization is a slightly more sophisticated approach that tries to return the words most basic dictionary form, or the lemma. Lemmatization might classify variations of *employment* differently dependent on the word's usage as a verb or noun. Lemmatization's' primary drawback is that it requires an external dictionary of lemmas (Manning et al. 2008).

Porter's (1980) stemming algorithm is the most common method of reducing inflectional complexity (Manning et al. 2008), although other algorithms also perform well. Stemming tends to increase recall, but reduces the precision of estimates as it treats some words with many variations as the same root (Manning et al. 2008). Stemming benefits some languages more than others, as linguistic morphology increases (in languages such as German or Finnish), stemming better accounts for the complexity. Stemming has been found to increase recall in a number of European languages, but does not consistently improve information retrieval for all cases (Hollink et al. 2004).

Following from this discussion, we propose that the removal of stopwords and stemming algorithms will increase the substantive differences between actors and reduce estimates of uncertainty for each speaker. As Hollink et al. (2004) find, however, the

improvements may not be comparable across all languages. Taking the cue from this research, in the next section, we describe a new data set on debates held at parties' national meetings and political leader speeches from four European democracies that will be used to assess the advantages and disadvantages of pre-processing across languages.

Methods and Techniques

Using *Wordfish*, a well-known unsupervised technique that performs scaling of political texts to extract the policy position of political actors who delivered those texts, we assess whether and to what extent pre-processing political texts affects the estimates of policy positions and the uncertainty around these estimates. We perform a comparative analysis focusing our attention on four different cases. These cases present both similarities and differences. In particular, we will analyze political texts belonging to four different countries (Denmark, Germany, Italy and the Netherlands) and written in four different languages, three of which are Germanic (Danish, Dutch, German), while one is Romanic (Italian).

For the reasons discussed above, and in particular for the close link between language and ideology, which is a feature required by (supervised and) unsupervised scaling techniques, we primarily analyze data from intra-party congress debates that are assumed to involve a higher degree of ideology given that the audience is composed by the party leadership and party activists. Here we investigate pre-processing distinguishing between oral speeches delivered by party leaders (in the German and Dutch case) and written motions provided by party faction (based on Italian data). However, going beyond the realm of intra-party politics, we will also evaluate whether our findings hold true for other types of political texts, such as the speeches delivered by Danish Prime Ministers.

We primarily look to texts produced at intra-party conference, because studies of intra-party politics have found that documents and speeches from parties' national meetings reveal important information about internal actors' behaviors. Furthermore, internal debates involving the party leadership and party activists tend to focus on policy issues and to adopt a language markedly characterized by references to ideology. In this regard, they neatly match

the 'ideological dominance' assumption required by unsupervised scaling of political texts (Grimmer and Stewart 2013).

While studies traditionally theorized that intra-party groups hold influence over leadership selection and behavior in parliament, tracing such behavior has been historically difficult. Improvements in automated text analysis have proved crucial in empirically understanding these relationships. For example, Bäck, Debus and Klüver (2014) link the manifestos of state level parties to federal manifestos in Germany. Ceron (2012, 2013 and 2014) demonstrates that motions given at party meetings in Italy can be used to locate the positions of intra-party factions and predict parties' behavior in government. Greene and Haber (2014) find through speeches at party meetings in France and Germany that intra-party disagreement increases when parties are in government, but expect to lose an election, though the location of party leadership candidates' revealed preferences in parliament are only important for parties in the opposition.

The speeches delivered by the Danish Prime Ministers are somewhat different compared to intra-party texts. Every year, these speeches mark the start of the new parliamentary session (held on the first Tuesday of October), and they are delivered in the Parliament, so the audience is broader compared to intra-party speeches. In recent years (but not in early periods), the media also covers these speeches extensively; this feature can contribute to changes in how the speeches are approached by the PM. Their importance, however, goes beyond simple formalities. As described in §38 (1) of the Danish Constitutional Act, these non-technical political speeches should offer an account of the current state of Danish affairs, and the speech is followed by (starting on Thursday) a lengthy debate on the opening address and government's financial plan, all parliamentary groups participating, with party spokespersons leading the debate. Overall, these speeches make references to previous achievements, but mostly set governmental priorities for the next parliamentary session, hence they are overarching, covering a mixture of topics, but focusing on the most salient few.

Accordingly, also driven by the time span covered, there is quite substantial heterogeneity in terms content, with the 'ideological dominance' being less pronounced compared to intra-party speeches. In this sense, these speeches serve the purpose of

evaluating whether our findings hold in general terms, even for political speeches that have different goals and target a slightly different audience than colleagues from one's party. Table 1 summarizes the number, source and period of these texts.

Table 1: Overview of speeches used in analysis

Country	**Period**	**Number of speeches**	**Source**	**Number of parties**
Denmark	1953-2013	61	PM speeches in parliament	4 (A, RV, V, C)
Germany	1990-2012	1660	Party leader speeches at party conferences	1 (CDU)
Italy	1989-2010	104	Written motions from party conferences	15 (Several parties)
Netherlands	1946-2013	126	Party leader speeches at party conferences	6 (VVD, CDA, PvdA, D66, GL, SP)

Stemming and Stopword Removal

To start with, our analyses consider the effect of excluding stopwords, using stemming or both, and contrast these results with those obtained when running *Wordfish* on word frequency matrices generated without any form of pre-processing. Table 2 provides details on the number of unique words considered in each analysis and the on the type of pre-processing technique adopted (if any) employed.

Table 2: Number of (unique) words pre- and post-processing

	Number of words (stems)			
	No processing	Stopword removal	Stemming	Stemming & Stopword removal
Denmark	20,429	20,355	13,213	13,187

Germany	57,686	57,324	41,360	41,158
Italy	38,864	38,656	20,634	20,512
Netherlands	27,627	27,547	23,040	23,020

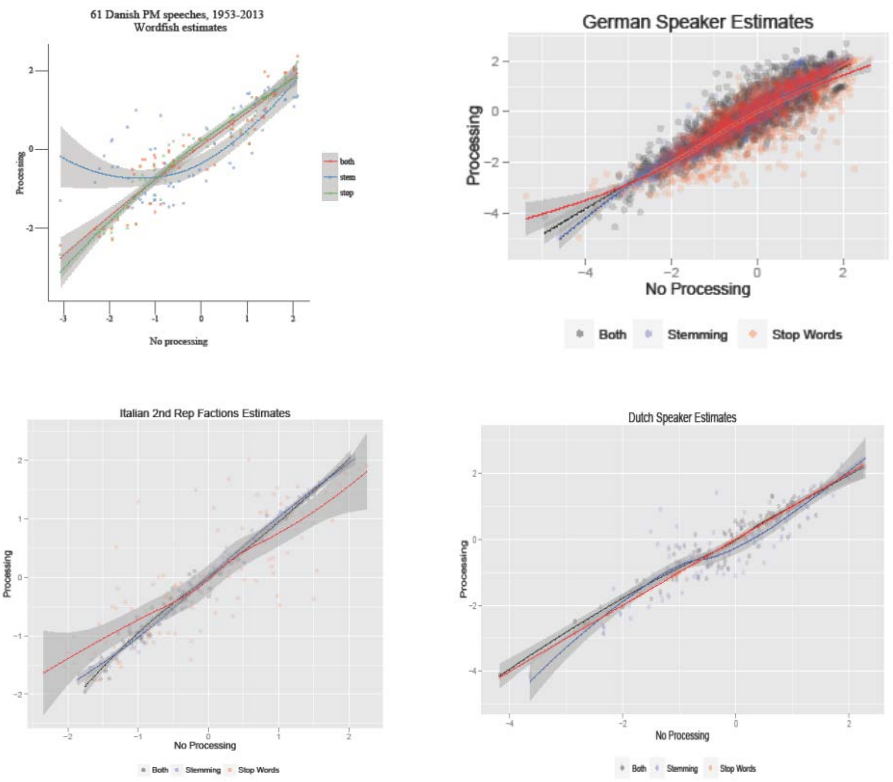
First, we look at the differences between the estimates. Figure 1 plots the different estimates using as reference (*x-axis*) the location retrieved from non-processed texts while positions after pre-processing (stopwords, stem, or both) are shown on the *y-axis*, for each country corpus.

Overall, the different measures seem to be highly correlated, but the picture also shows that pre-processing can sometimes make the difference. In the Dutch and the Italian case, the estimates appear to be quite similar to each other as the dots are quite close to the bisecting line. In Dutch texts, this is particularly true when we only remove stopwords, while there are more changes when using stemming (especially for ideological 'moderate' documents). In the Italian case, to the contrary, it is removing stopwords that makes the difference while the estimates of stemming and no-processing are highly correlated (0.99).[2]

In the German and Danish texts, dots are scattered further away from the bisecting line. Here we observe, once again, that stopwords and stemming does not seem to matter in the same way. In Germany, the results look quite different when we remove stopwords, while in the Danish case stemming seems more an issue, particularly for speeches originally on the lower end of the latent dimension scale.

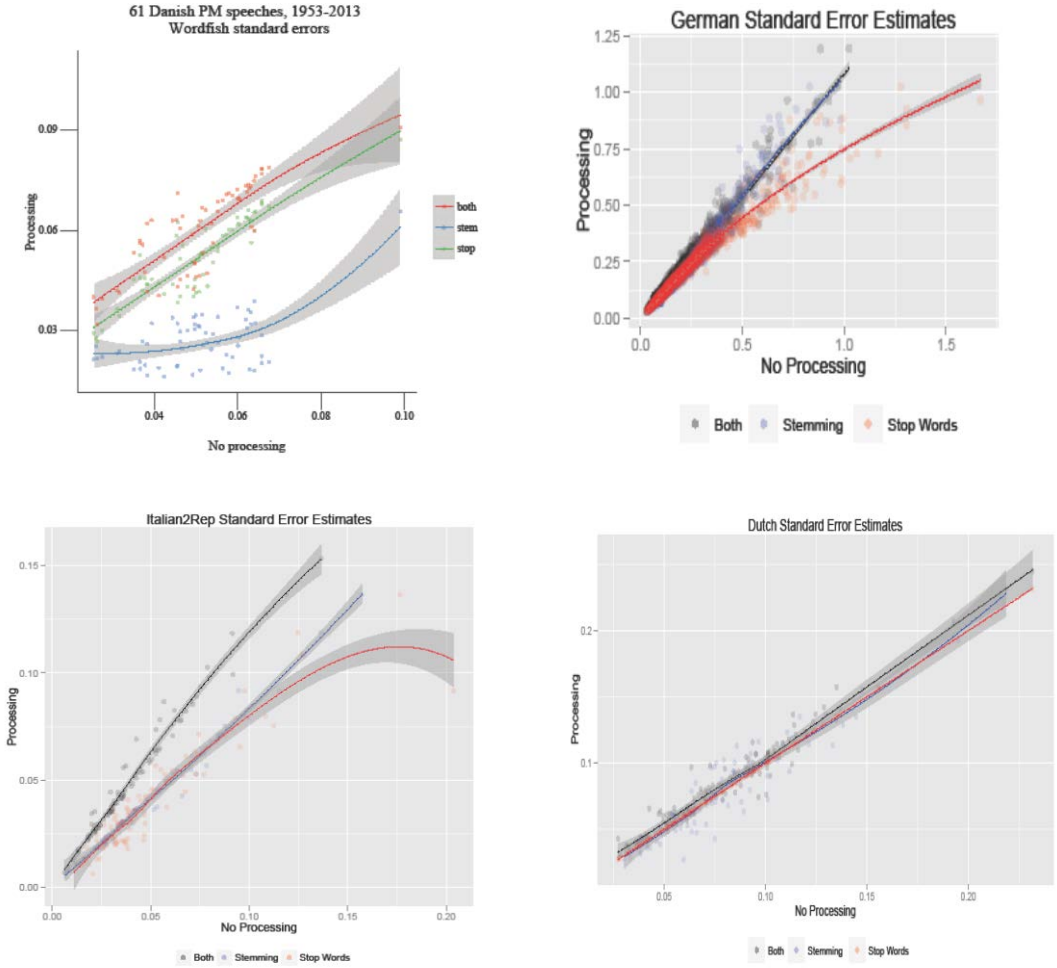
[2] One explanation might be the length of the list of stopwords of each language. The Dutch list has approximately 90 words, the Italian approximately 200, the German 300, and the Danish 80.

Figure 1: Speaker estimates after stopword removal, stemming, or both



We focus on what happens to the standard errors of the estimates in Figure 2. Here the differences become wider, except for the Dutch case in which we do not observe many differences. Overall, removing stopwords also reduces the magnitude of the standard errors (both in Germany, Italy and Denmark). Conversely, using the stem of words is effective in reducing standard errors in Danish and Italian texts, while it has a negative effect in German speeches. Remarkably, combining stemming and stop words yields negative consequences on the uncertainty of the estimates: in all the four countries standard errors tend, in fact, to increase. Why do stopword removal and stemming, taken alone, reduce uncertainty, while when both are combined the standard errors start to grow? This perhaps surprising finding suggests that scholars have to weigh the effect of different forms of pre-processing before running the analysis and highlights the need to investigate more in depth this matter.

Figure 2: Standard errors after stopword removal, stemming, or both



Removal of sparse words

In the next step, we remove sparse terms. For this purpose we use the term-document matrices that have been stemmed and from which stopwords have been removed (see rightmost column in Table 2). Subsequently, we remove terms that occur in fewer than respectively 1%, 5%, 10% and 20% of the documents. Table 3 displays the number of unique words that remain in the term-document matrix after excluding the sparse terms. The number of words that remain differs strikingly between languages. From the German data only 0.4% remain if we remove terms that occur in fewer than 20% of the documents. For the Dutch data 3.5% remain, and for

the Italian and Danish data around 10% remain. Especially, the difference between linguistically close languages such as Dutch and German is striking.

Table 3: Removal of sparse words

	No removal*	<1%	<5%	<10%	<20%
Denmark	13,187	13,187	3,441	2,227	1,327
Germany	41,158	3,577	911	407	176
Italy	20,512	10,770	5,136	3,355	2,065
Netherlands	23,020	8,613	2,778	1,603	812

* Number of words left over after stemming and stopword removal.

In Figure 3 we compare the position of a document where no sparse terms have been removed to the positions of the same document where we did remove sparse terms.[3] The pictures reveal a number of unexpected trends. For the Danish, German and Dutch cases all evidence a non-linear relationship between the estimates with sparse word removal and those without removing sparse words. The extent of the non-linearity becomes more pronounced as the number of sparse words is increased. Furthermore, the non-linearity seems to be focused on the negative end of the scale. This suggests that the specific words removed were given some weight to distinguish one end of the scale, but were then removed as sparse words.

The correlations between the estimates after removing sparse words are remarkably low, with the exception of the Italian case. For example, in the Netherlands these correlations are shockingly low. The correlation between documents where no sparse terms have been removed, and documents where only 1% was removed is 0.62. If you compare the former to documents where 10% or 20% were removed it is only 0.14. This suggests that researchers should remove sparse terms with extreme care, also to avoid losing many informative words.

Figure 4 illustrates the standard error estimates for each country as sparse words are removed. For each case, the removal of greater sparse words tends to decrease the size of the standard errors. The removal of too many words may improve the model's ability to distinguish

[3] Results for 1% sparsity in Denmark not displayed, as they are identical to those with no sparse term reduction.

between texts, but the reduced standard errors also might lead to overconfidence in the derived estimates.

Figure 3: Speaker estimates after sparse word removal

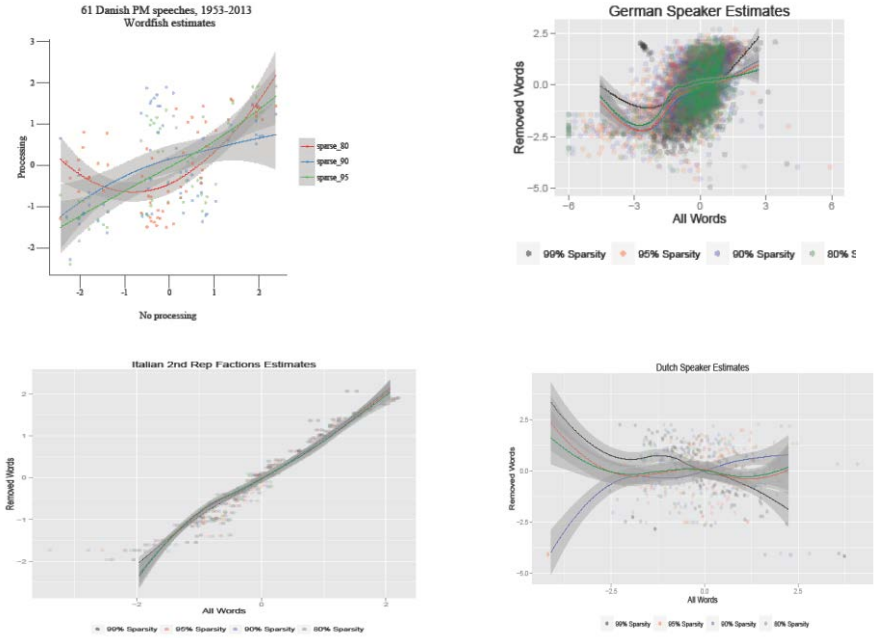
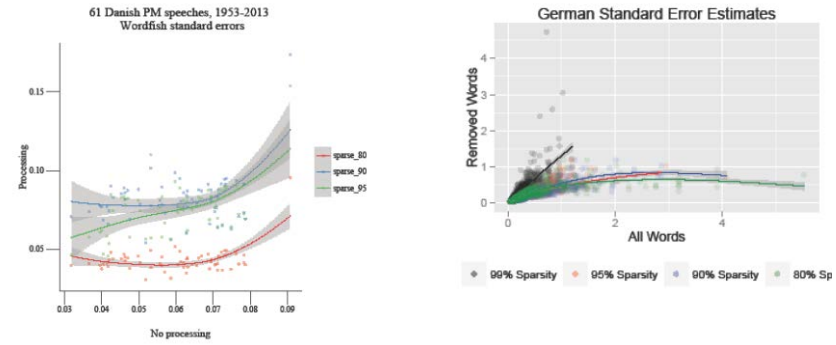
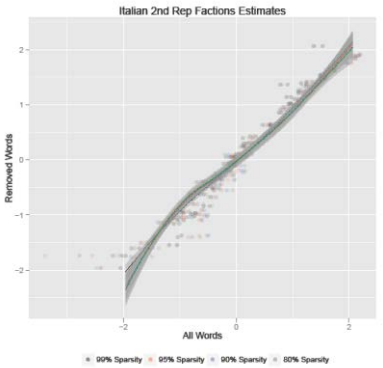
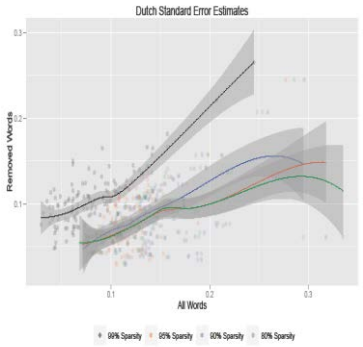


Figure 4: Standard error estimates after sparse word removal







Analyzing Distance in Position and Standard Errors Including Covariates

In a next step, we analyze the absolute distances in the position of speaker estimates produced by the various methods compared to the position of speaker estimates of unprocessed texts. We use the pre-processing methods as independent variables and in two cases add covariates such as time and party to a standard OLS regression analysis. Table 4 displays the results.

Table 4: Regression results (OLS) for distance to position retrieved from unprocessed document

	Netherlands	Germany	Denmark	Italy
Stemming	-0.007	-0.003	0.084	-0.413***
	(0.105)	(0.021)	(0.102)	(0.033)
Stopwords	-0.029	0.222***	-0.334***	0.019
	(0.105)	(0.021)	(0.102)	(0.033)
Removed 1% sparse terms	1.479***	0.289***	0.000	0.068**
	(0.105)	(0.021)	(0.102)	(0.033)
Removed 5% sparse terms	0.861***	0.424***	0.596***	0.130***
	(0.105)	(0.021)	(0.102)	(0.033)
Removed 10% sparse terms	1.251***	0.482***	0.911***	0.157***
	(0.105)	(0.021)	(0.102)	(0.033)
Removed 20% sparse terms	1.201***	0.446***	0.327***	0.171***

	(0.105)	(0.021)	(0.102)	(0.033)
Before 1970	0.129		0.301***	
	(0.098)		(0.077)	
Between 1970 and 1990	0.104		-0.014	
	(0.075)		(0.073)	
Party 1[4]	0.056		-0.562***	
	(0.102)		(0.128)	
Party 2	0.512***		-0.136*	
	(0.114)		(0.081)	
Party 3	0.118		-0.421***	
	(0.097)		(0.103)	
Party 4	0.175			
	(0.130)			
Party 5	0.400***			
	(0.103)			
Constant	-0.008	0.227***	0.743***	0.489***
	(0.147)	(0.025)	(0.147)	(0.041)
Observations	882	11,620	427	728
Adjusted R^2	0.368	0.155	0.322	0.234
Note:		*p < 0.1 **p< 0.05 ***p< 0.01		

Again, we find that stemming and stopword removal have different effects across languages: no effect in Dutch, a positive effect of stopword removal in German, a negative effect of stopword removal in Denmark, and a negative effect of stemming in Italy. Removing sparse terms has a similar effect across languages: it increases the distance to the unprocessed

[4] Parties 1 to 5 in the Netherlands are D66, GL, PvdA, SP and VVD. In Denmark parties 1 to 3 are RV, A and V.

document. The covariates matter too. The Danish and Dutch models have a much higher explained variance than the German and Italian models.[5] Estimates of some parties in some time periods are apparently more vulnerable to model specification.

Table 5: Regression results (OLS) for distance to SE retrieved from unprocessed document

	Netherlands	Germany	Denmark	Italy
Stemming	-0.005	-0.006	-0.008***	0.001
	(0.003)	(0.006)	(0.002)	(0.002)
Stopwords	-0.005	-0.013**	-0.033***	-0.001
	(0.003)	(0.006)	(0.002)	(0.002)
Removed 1% sparse terms	-0.023***	-0.073***	0.000	-0.005***
	(0.003)	(0.006)	(0.002)	(0.002)
Removed 5% sparse terms	-0.042***	-0.127***	-0.016***	-0.011***
	(0.003)	(0.006)	(0.002)	(0.002)
Removed 10% sparse terms	-0.057***	-0.173***	-0.022***	-0.015***
	(0.003)	(0.006)	(0.002)	(0.002)
Removed 20% sparse terms	-0.057***	-0.122***	0.017***	-0.020***
	(0.003)	(0.006)	(0.002)	(0.002)
Before 1970	-0.002		0.011***	
	(0.003)		(0.001)	
Between 1970 and 1990	-0.006***		0.003***	
	(0.002)		(0.001)	
Party 1	-0.002		-0.001	
	(0.003)		(0.002)	
Party 2	-0.007*		0.0004	

[5] This could be because of the longer time component.

	(0.003)		(0.001)	
Party 3	0.005*		-0.001	
	(0.003)		(0.002)	
Party 4	0.007*			
	(0.004)			
Party 5	0.005			
	(0.003)			
Constant	0.001	0.001	0.028***	-0.009***
	(0.004)	(0.008)	(0.002)	(0.002)
Observations	882	11,620	427	728
Adjusted R^2	0.541	0.130	0.786	0.234
Note:				*p<0.1 **p<0.5 ***p<0.01

We perform a similar analysis on the standard errors produced by the different models. However, this time we are interest in relative differences with estimates from the unprocessed model, because we want to analyze whether pre-processing decisions increase or decrease error. Table 5 displays these results. Most processing procedures reduce error, although – especially in the case of stemming – this effect is not always significant. The effect of removing sparse terms is almost always statistically significant and, with the exception of Denmark, it is much larger than that of stemming and removing stopwords. Similar to the previous analyses the models with covariates perform much better in terms of explained variance. The effects of the party and time dummies indicate that some parties and some time periods errors are systematically larger or smaller.

Conclusions

Automated text analysis holds great potential for understanding political behavior. Evidence on intra-party politics and government behavior has already revealed substantial novel insights. Much attention has already been given to the choice between types of text analytic models in political science. So far, however, less attention has been paid to the methods for dealing with linguistic complexity prior to the analysis stage. Our results show that common methods for preparing text for analysis have serious implications for the estimates derived using the *Wordfish* algorithm.

The differences do not vary consistently across languages. We caution researchers to remove stopwords and to stem documents, as the outcome of this choice seems case-dependent. To the contrary, the implications of removing sparse words seem to be similar across languages. Indeed, removing sparse words leads to large differences in the estimated speaker positions and reduces the uncertainty of the estimates in nearly every case. Notice that sparse word removal may cause unexpected shifts in speakers' positions and suggest greater certainty than is warranted. Furthermore, the removal of sparse words should not be taken lightly as words occurring in only a small number of documents are likely to be given low word weights and have little impact on the final estimates to begin with. We suggest that scholars be transparent about the other pre-processing strategies undertaken and pay close attention to differences based on preprocessing decisions rather than theoretically derived differences in the texts.

References

Bäck, Hanna, Marc Debus, and Heike Klüver. 2014. "Bicameralism, Intra-Party Bargaining, and the Formation of Party Policy Positions Evidence from the German Federal System." *Party Politics*: 1354068814549343.

Baerg, Nicole Rae. 2014. "War of the Words: How Elites' Communication Changes the Economy." http://mpra.ub.uni-muenchen.de/59823/ (February 19, 2015).

Baerg, Nicole Rae, and Will Lowe. 2015. "Estimating Central Bank Preferences Combining Topic and Scaling Methods." http://mpra.ub.uni-muenchen.de/61534/ (February 19, 2015).

Beauchamp, Nick. 2011. *Using text to scale legislatures with uninformative voting*. New York University Mimeo.

Benoit, Kenneth, and Alexander Herzog. 2015. "Text Analysis: Estimating Policy Preferences From Written and Spoken Words." In: Jennifer Bachner, Kathyrn Wagner Hill, and Benjamin Ginsberg (eds.) *Analytics, Policy and Governance*, forthcoming.

Ceron, Andrea. 2012. "Bounded Oligarchy: How and When Factions Constrain Leaders in Party Position-Taking." *Electoral Studies* 31(4): 689–701.

———. 2013. "Brave Rebels Stay Home: Assessing the Effect of Intra-Party Ideological Heterogeneity and Party Whip on Roll-Call Votes." *Party Politics*: 1354068812472581.

———. 2014. "Inter-Factional Conflicts and Government Formation Do Party Leaders Sort out Ideological Heterogeneity?" *Party Politics*: 1354068814563974.

Coscia, Michele, and Viridiana Rios. 2012. "Knowing Where and How Criminal Organizations Operate Using Web Content." In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, ACM, 1412–21. http://dl.acm.org/citation.cfm?id=2398446 (February 25, 2015).

Denny, Matthew, and Arthur Spring. 2016. "Assessing the Consequences of Text Preprocessing Decisions." *SSRN* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2849145 (Accessed October 26).

Eggers, Andrew C., and Arthur Spirling. 2014. "Party Cohesion in Westminster Systems: Inducements, Replacement and Discipline in the House of Commons, 1836–1910." *British Journal of Political Science*: 1–23.

Giannetti, Daniela, and Michael Laver. 2005. “Policy positions and jobs in the government.” *European Journal of Political Research* 44(1): 91–120.

Greene, Zachary, and Matthias Haber. 2014. “Leadership Competition and Disagreement at Party National Congresses.” *British Journal of Political Science* FirstView: 1–22.

Grimmer, Justin. 2010. “A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases.” *Political Analysis* 18(1): 1–35.

Grimmer, Justin, and Brandon M. Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis*: mps028.

Hollink, Vera, Jaap Kamps, Christof Monz, and Maarten de Rijke. 2004. “Monolingual Document Retrieval for European Languages.” *Information Retrieval* 7(1-2): 33–52.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. “How Censorship in China Allows Government Criticism but Silences Collective Expression.” *American Political Science Review* 107(02): 326–43.

Klemmensen, Robert, Sara Binzer Hobolt, and Martin Ejnar Hansen. 2007. “Estimating Policy Positions Using Political Texts: An Evaluation of the Wordscores Approach.” *Electoral Studies* 26(4): 746–55.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. “Extracting Policy Positions from Political Texts Using Words as Data.” *American Political Science Review* null(02): 311–31.

Loewenberg, Gerhard. 2008. “The Contribution of Comparative Research to Measuring the Policy Preferences of Legislators.” *Legislative Studies Quarterly* 33(4): 501–10.

Lo, James, Sven-Oliver Proksch, and Thomas Gschwend. 2014. “A Common Left-Right Scale for Voters and Parties in Europe.” *Political Analysis* 22(2): 205–23.

Lowe, Will. 2008. “Understanding Wordscores.” *Political Analysis* 16(4): 356–71.

Lucas, Christopher et al. 2015. “Computer-Assisted Text Analysis for Comparative Politics.” *Political Analysis*: mpu019.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. “Fightin’words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict.” *Political Analysis* 16(4): 372–403.

Proksch, Sven-Oliver, and Jonathan B. Slapin. 2009a. “How to Avoid Pitfalls in Statistical Analysis of Political Texts: The Case of Germany.” *German Politics* 18(3): 323–44.

———. 2009b. WORDFISH: Scaling Software for Estimating Political Positions from Texts.

———. 2012. “Institutional Foundations of Legislative Speech.” *American Journal of Political Science* 56(3): 520–37.

Quinn, Kevin M. et al. 2010. “How to Analyze Political Attention with Minimal Assumptions and Costs.” *American Journal of Political Science* 54(1): 209–28.

Schonhardt-Bailey, Cheryl. 2005. “Measuring Ideas More Effectively: An Analysis of Bush and Kerry’s National Security Speeches.” *Political Science and Politics* 38(04): 701–11.

Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. “A Scaling Model for Estimating Time-Series Party Positions from Texts.” *American Journal of Political Science* 52(3): 705–22.

Spirling, Arthur. 2012. “US Treaty Making with American Indians: Institutional Change and Relative Power, 1784–1911.” *American Journal of Political Science* 56(1): 84–97.

Stockmann, Daniela. 2012. *Media Commercialization and Authoritarian Rule in China*. Cambridge University Press.