PA

Testing the Validity of Automatic Speech Recognition for Political Text Analysis

Sven-Oliver Proksch[1], **Christopher Wratil**[2,1] **and Jens Wäckerle**[1]

[1] *Cologne Center for Comparative Politics, University of Cologne, Germany. Email: so.proksch@uni-koeln.de*
[2] *Minda de Gunzburg Center for European Studies, Harvard University, Cambridge, MA 02138, USA*

Abstract

The analysis of political texts from parliamentary speeches, party manifestos, social media, or press releases forms the basis of major and growing fields in political science, not least since advances in "text-as-data" methods have rendered the analysis of large text corpora straightforward. However, a lot of sources of political speech are not regularly transcribed, and their on-demand transcription by humans is prohibitively expensive for research purposes. This class includes political speech in certain legislatures, during political party conferences as well as television interviews and talk shows. We showcase how scholars can use automatic speech recognition systems to analyze such speech with quantitative text analysis models of the "bag-of-words" variety. To probe results for robustness to transcription error, we present an original "word error rate simulation" (WERSIM) procedure implemented in *R*. We demonstrate the potential of automatic speech recognition to address open questions in political science with two substantive applications and discuss its limitations and practical challenges.

Keywords: Google, YouTube, text analysis, transcriptions, automatic speech recognition, campaigns

1 Introduction

Despite the wide availability of transcribed political speech, including parliamentary debates, party press releases, or legal decisions, numerous political debates still escape scholarly analysis due to the unavailability of speech transcriptions. This prevents, or at least seriously impedes, systematic analysis of many informal political arenas. For example, politicians regularly participate in talk shows and interviews on radio or television. Such outlets are interesting from a research perspective, as they allow politicians to speak more freely than in a parliamentary arena, where speaking time is limited and party discipline dominates (Proksch and Slapin 2012, 2015; Herzog and Benoit 2015). However, most research on political talk shows and interviews focuses on their effects on voters rather than the actual content of political speeches (see e.g., Baum 2005; Baum and Jamison 2006).

Political parties and their candidates also increasingly communicate directly with constituents by maintaining their own channels on the YouTube video portal (see e.g., Haynes and Pitts 2009; Gibson and McAllister 2011; Zittel 2015). In addition to what members post on social media, these YouTube sites often bundle political speeches of party members at public events or intra-party conferences, providing a rich source of information about developments inside political parties. But virtually no work has comprehensively analyzed the content of such channels.

Political Analysis (2019)
vol. 27:339–359
DOI: 10.1017/pan.2018.62

Published
19 February 2019

Corresponding author
Sven-Oliver Proksch

Edited by
Jeff Gill

Authors' note: We are grateful to Leonie Diffené, Felix Reich and Pit Rieger for their excellent research assistance. We are also very thankful for helpful comments on earlier versions of this work by two anonymous reviewers as well as Jeff Gill. Christopher Wratil would like to acknowledge funding by the Fritz Thyssen Stiftung (20.16.0.045WW). All remaining errors are our own. The replication files for this article are available on the *Political Analysis* Dataverse (Proksch, Wratil, and Wäckerle 2018).

Yet another example in which spontaneous and nontranscribed speech may play an important role are political protests and demonstrations, such as the "Occupy" movement in the U.S., the "Indignados" protests in Spain, and "Pegida" demonstrations in Germany. These events are organized on the streets with participants delivering speeches in front of a crowd of people during the protests. Oftentimes, videos of these protests are available, thus providing an additional source of information about how social movements develop. But even the more recent literature on protest events has virtually neglected videos of protest speeches as an analysis source (e.g., Kriesi *et al.* 2012; Accornero and Ramos Pinto 2015; Giugni and Grasso 2015; Dolezal, Hutter, and Becker 2016). Finally, in international relations, governments frequently negotiate with each other and some of these negotiations are recorded on video but not transcribed, as in some international climate change negotiations or the Council of the European Union (McKibben 2016; Wratil and Hobolt Forthcoming). Such recordings may provide novel insights into how intergovernmental negotiations unfold.

In short, there is an abundance of nontranscribed recordings of political speech which largely escapes scholarly analysis despite its potential importance in the political process. Automatic speech recognition (ASR) technology may provide an efficient way for generating text corpora of such speeches. In this study, we are examining the potential and limitations of text generated via ASR for political text analysis. Specifically, we use Google's leading speech recognition technologies accessible through YouTube and the Google Speech API. If textual quantities of interest from such corpora have a high validity, then this opens up a whole new range of hitherto unexplored text sources that political scientists can investigate systematically. Recent methodological advances in political speech analysis have focused on audio or video recordings, but not for the purpose of establishing whether automatic transcriptions are valid. For instance, Knox and Lucas (2018) as well as Dietrich et al. (2017, 2018) analyze the audio signal of political speech as data to extract features such as emotional arousal, which transcriptions alone fail to reflect. In this study, we demonstrate that ASR technology can indeed be used to systematically analyze untapped recordings of political speech with quantitative text-analytic approaches (e.g., Grimmer and Stewart 2013).

Specifically, we validate the use of automatic transcriptions for the Wordfish scaling model (Slapin and Proksch 2008) as well as dictionary-based approaches to sentiment (e.g., Young and Soroka 2012; Daku, Soroka, and Young 2015; Proksch *et al.* 2018) as two prominent "bag-of-words" models using a corpus of "State of the Union" debates from the European Union. To assess the sensitivity of substantive findings to transcription errors stemming from ASR we present a "word error rate simulation" (WERSIM) procedure, implemented in an *R* package that researchers can use and adapt for any text-analytic project they pursue. To illustrate the potential of ASR, we apply scaling methods to two new text corpora generated via ASR: a campaign speech corpus for a parliamentary election and an intergovernmental negotiations text corpus from the European Union. In both settings, we are able to gain new insights into the political process that would otherwise be difficult or impossible. We conclude with several best-practice recommendations for implementing ASR in political science research designs.

2 Automatic Speech Recognition Systems for Political Speech

Initial attempts at ASR date back to the 1950s, but it took until the early 1990s for the first commercial products to enter the consumer markets. While traditional ASR architectures had combined hidden Markov models with Gaussian mixture models, major improvements in accuracy were achieved through the use of deep neural networks with many hidden layers, which can be effectively trained on a large amount of data given advances in hardware, e.g., "graphics processing units" (for an overview see, e.g., Hinton *et al.* 2012; Lecun, Bengio, and Hinton 2015; Yu and Deng 2015).

A distinguishing feature of ASR systems is the level of their *speaker dependence*. Whereas speaker-dependent systems are trained to recognize speech from the voice of a particular individual, speaker-independent systems attempt to deal with any speaker. This distinction is highly relevant, since a lot of political speech is inherently produced by multiple speakers (e.g., political discussions, negotiations, parliamentary debates), and training the system to each speaker's voice is rather costly. Hence, for most applications in political science, only *speaker-independent* ASR systems have the potential to radically simplify the transcription of speech.

During the last years, in particular, technology companies like Google and Microsoft have developed their own speaker-independent ASR systems, which they use in their own products and also make available to customers through websites (e.g., YouTube) or an application programming interface (API). These systems are often based on recurrent neural networks, such as long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997), and many improve quickly over time by continuously learning from user contributions (e.g., user-provided captions for videos on YouTube) (e.g., Liao, McDermott, and Senior 2013; Soltau, Liao, and Sak 2017). In addition, open-source systems for multiple speakers have been developed, such as "Sphinx 4".

While our interest in this study is to test the validity of ASR for political text analysis, it is important to note that developers of ASR systems typically focus on accuracy. To the best of our knowledge, Google's ASR systems have reached the highest accuracy of the prominent competition (see also Këpuska and Bohouta 2017). The most commonly used measure to assess accuracy is the *word error rate* (WER), which compares a hypothesis text (the ASR-transcribed text) with a reference text (e.g., the text that was read out to the ASR system). It is defined as:

$$WER = \frac{S + D + I}{N}. \tag{1}$$

Where N is the number of words in the reference text, S is the number of substituted words in the hypothesis text (i.e., words with inaccurate ASR transcription), D is the number of deletions in the hypothesis text (i.e., words that are missing in the ASR transcription), and I is the number of insertions in the hypothesis text (i.e., words in the ASR transcription that are not in the reference text). Words are only classified as correct, if the word in the hypothesis text *exactly* matches the word in the reference text. According to its own assessment, Google has reached WERs around 0.05 for English language texts by summer 2017, rapidly bringing rates down from >0.2 back in 2013 (Meeker 2017).

However, there are concerns that the WER may be misleading. On the one hand, indicated high levels of accuracy are obtained under laboratory conditions and WERs for typical sources of political speech (e.g., including background noise, speaker interference) may therefore be higher. On the other hand, we do not have any benchmark for how low WERs must be in order to be acceptable for typical political science applications. Ultimately, political scientists do not care about the actual WER but whether the final *quantities of interest* they retrieve from political texts (e.g., position estimates, word/dictionary counts, topic distributions) are sufficiently similar when speech is ASR-transcribed as opposed to human-transcribed. The mapping, however, between quantities of interest and WERs is unknown and may vary from corpus to corpus and between analysis models. In the following sections, we assess the validity of ASR for political text analysis by comparing human and ASR-based transcriptions as well as by applying the WERSIM procedure to test the robustness of results to typical ASR transcription errors.

3 Validating ASR with the SOTEU Corpus

To systematically compare final quantities of interest when political speech is transcribed by ASR versus by humans we use speeches in the European Parliament's (EP) annual "State of the Union"

(SOTEU) plenary session.[1] Similar to the "State of the Union Address" given by the U.S. president, the president of the European Commission addresses the members of the European Parliament (MEPs) with her view on the state of political affairs in the EU. In contrast to the U.S. model, a full parliamentary debate follows in the EP, with MEPs from all party groups responding to the president's speech. Since 2010, SOTEU debates are one of the most visible sessions of the EP. Importantly, since all speech is simultaneously interpreted into all official EU languages, SOTEU provides us with the unique opportunity to test ASR across several languages holding the content of the political speeches constant. Therefore, we can make accurate comparisons of the validity of ASR across languages. Specifically, we obtain recordings of the 2011 SOTEU session with an audio track of the simultaneous translation (where applicable) in English, French and German. In total, the 2011 SOTEU session contained speech interventions by 57 speakers, for whom we pool all interventions by a speaker into a single text document.[2]

We obtain two sets of ASR transcriptions of these video recordings. First, we use a transcription of the videos produced by the Google Speech API. Second, we use a transcription based on automatic captioning of videos by Google's YouTube video portal. Below we show that there are nontrivial differences in the quality of the transcriptions produced by the API and YouTube, with YouTube reaching higher accuracy. All our ASR transcriptions were produced in spring 2018. The ASR transcriptions serve as two forms of hypothesis texts. We use human transcriptions of the sessions as our reference texts or "gold standard". Research assistants with proficiency in the respective language produced "word-by-word" transcriptions of the videos, including word repetitions and slips of the tongue that would also be transcribed by an ASR system.[3]

3.1 Comparing Word Accuracy

We first assess the accuracy and the similarity of the Google transcriptions with our human transcriptions. The average WER is 0.03 for a speech in English, 0.12 for one in French and 0.10 for one in German when the YouTube transcriptions are used, and 0.21, 0.26 and 0.21 when the API transcriptions are used. The rates for YouTube are actually below Google's proclaimed level of about 0.05 for English. We also assess the cosine similarity of the transcriptions across speakers (see de Vries, Schoonvelde, and Schumacher 2018). The average similarities are 0.99 for English, 0.97 for French and 0.97 for German texts. For the API, the similarities are 0.96 for English, 0.94 for French and 0.94 for German. Details are contained in Appendices 1 and 2.

In the next step, we explore the components of the WER in our corpora. Figure 1 shows the distribution of deletions (D), insertions (I) and substitutions (S). In all corpora, substitutions make up the biggest share of mistakes by the ASR system. Except for the English YouTube corpus, deletions constitute the second highest share of errors. They are especially prevalent in the API transcriptions, as the ASR of the Google API more often does not return any word, while YouTube tends to return a suggestion. This suggestion might either be correct, and reduce the error, or incorrect, leading to a substitution. Insertions only play a minor role. Hence, these results suggest that YouTube is particularly outperforming the Google API on deletions and substitutions. In general, these figures indicate a rather high accuracy and similarity of ASR transcriptions of political speech with human transcriptions. But the pertinent question is whether this translates into high similarities in quantities of interest from text models.

1 See Proksch, Wratil, and Wäckerle (2018) for the replication materials.
2 We exclude the president of the EP from all estimations, since she has an entirely nonpolitical role as moderator of the session.
3 Details of the language proficiency of our assistants and transcription rules are in Appendix 18. We took guidance from the rules laid out in Dresing, Pehl, and Schmieder (2015).

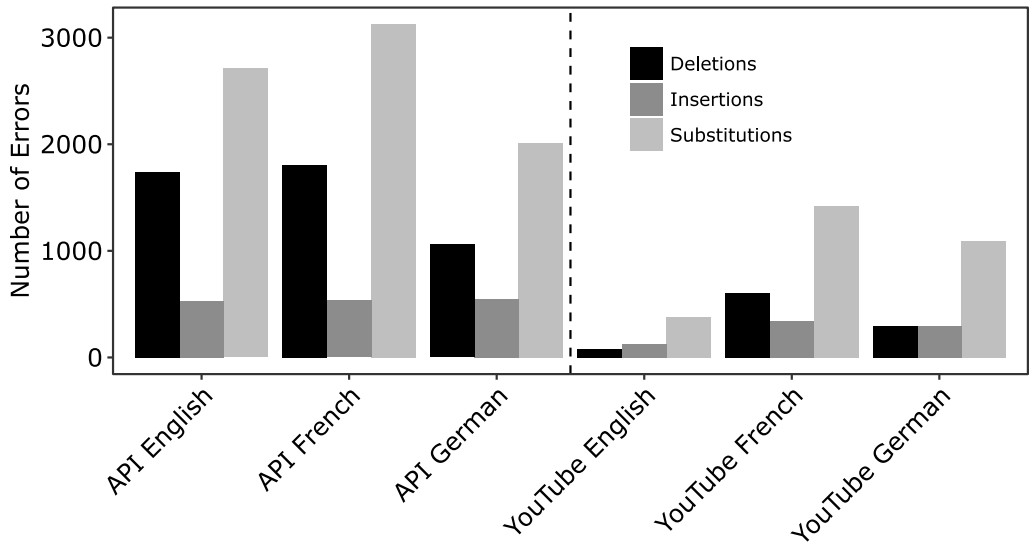

Figure 1. Word error structure in the State of the European Union debate.

3.2 Comparing Quantities of Interest

Next, we assess the cross-validity of quantities of interest from "bag-of-words" models. We cover two major classes of models, namely "scaling" and "classification" models (see Grimmer and Stewart 2013, 268), and select one appropriate method for our corpus from each. First, we apply the text scaling model Wordfish (Slapin and Proksch 2008). Second, we also test dictionary-based approaches by employing an English sentiment dictionary by Young and Soroka (2012) and its translation into German and French provided by Proksch *et al.* (2018). We use "sentiment" as our quantity of interest calculated as the log ratio of positive to negative words used by a speaker.[4] One alternative classification method to dictionary-based models are topic models (e.g., Grimmer 2010; Roberts *et al.* 2014). However, since such models are usually estimated on corpora containing thousands of documents and a large number of latent topics, our corpus of 57 documents from a single debate is not a typical application for such models. In Appendix 7, we nevertheless test whether topic distributions from ASR transcriptions are different from those based on human transcriptions by dividing the SOTEU corpus into word chunks containing 20 tokens each. Subsequently, we include both corpora in a single structural topic model (Roberts *et al.* 2014) using the transcription mode (human versus YouTube) as a covariate. The results show no difference in topic proportions depending on the transcription mode (see Figure A11 in the Appendix).

We note that decisions about preprocessing may influence the similarity of the document-feature matrices of the ASR-based and the human-transcribed corpora, and hence, to what degree quantities of interest from text models will be similar (see also Denny and Spirling 2018). First, the *removal of stopwords* may make a difference, since stopwords are common words that ASR systems tend to be good at identifying correctly. As a consequence, the removal of stopwords may diminish the performance of ASR transcriptions, since the WER on the remaining words may be higher. Second, *stemming* may influence performance as it renders ASR errors on singular versus plural forms, declension, and grammatical conjugation (especially important in French and German) irrelevant. However, such mistakes may be random or, at least, uncorrelated with features driving the estimation of quantities of interest. Hence, we cannot tell *a priori* whether the mistakes will have an impact but instead assess the impact empirically using the SOTEU corpus.

4 In Appendix 8 we also demonstrate that ASR transcriptions perform almost equally well when analyzed with substantially shorter dictionaries than the employed sentiment dictionaries.

Third, the *removal of infrequent features* may make a difference as the performance of ASR may be weak on uncommon words. Moreover, erroneous insertions by ASR may be removed as infrequent features. Hence, we expect that the removal of infrequent words will improve the accuracy of quantities of interest derived from ASR transcriptions. For our corpus, we assess the consequences of removing features that occur less than five times in the corpus. Note that we cannot assess the consequences of removing punctuation, since at the time of writing Google's ASR system is not able to identify punctuation. We also do not assess the influence of number removal and lowercasing, since there is wide agreement that these preprocessing steps should be performed. In addition to preprocessing decisions, we also assess the influence of text length. A longer text does not lead to a higher accuracy of ASR transcriptions (i.e., lower WERs), but the influence of transcription errors on quantities of interest may simply cancel out as the number of words increases. Hence, we assess the improvement in our ASR-based estimates when we drop texts below 100 words as well as below 200 words.

Figure 2 displays the correlations between our quantities of interest when scaled from human "word-by-word" transcriptions versus ASR transcriptions. For the Wordfish model (Figure 2a), these are the correlations between the position estimates for the speakers (θ_i). For the dictionary-based analysis (Figure 2b), these are the correlations between the sentiment estimates for the speakers. In our "baseline" specification, we preprocess the texts with lowercasing, number and punctuation removal, and—only for Wordfish but not sentiment—a minimum feature frequency of three.[5] We then assess the influence of the other preprocessing steps separately by adding them to our baseline specification.

First, the results show that the ASR transcriptions by YouTube more accurately recover the quantities of interest obtained from human transcriptions than transcriptions by the Google API. The correlations between YouTube transcriptions and human transcriptions are always higher than (or equal to) those between API transcriptions and human transcriptions. Second, ASR transcriptions are generally doing well in recovering the quantities of interest from human transcriptions: in 30 out of 33 specifications (18 for Wordfish and 15 for sentiment), human- versus YouTube-based quantities of interest correlate at >0.9.[6] For English, the lowest correlation across all specifications is 0.96. Only in German do the removal of stopwords as well as the introduction of a minimum feature frequency of five words have substantial consequences for the correlation of the estimated Wordfish positions. Nevertheless, all correlations remain at >0.7. This confirms our conjecture that stopwords, as "easy" words, help to align ASR-based estimates with those based on human transcription (we observe a similar negative impact of stopwords removal in French but not in English). However, the negative influence of the removal of infrequent features runs counter to our expectations. In our limited corpus the removal of infrequent features drastically reduces the number of unique features in the corpus (from 2743 to 591), which seems to render estimates more fickle, depending on which exact feature is removed from the YouTube- versus human-transcribed corpus. Furthermore, we observe that stemming has some influence on the performance of ASR transcriptions, slightly improving results for the French Wordfish model but deteriorating them for the German language. This may suggest that the influence of stemming is rather language-dependent. Last, the results show, with the exception of the sentiment estimates in French, that the accuracy of ASR-based estimates is higher for longer speeches, in particular if texts shorter than 200 words are removed. In sum, preprocessing steps have little influence on the validity of ASR-based estimates when using transcriptions of the quality achieved by YouTube.[7]

5 Throughout the paper we implement all Wordfish models using the "austin" package in *R*. Note that we remove infrequent features (less than three occurrences across the corpus), as such features would obtain excessive word weights.

6 We do not assess the influence of stemming for the sentiment analysis, since sentiment dictionaries contain word stems with related wildcards.

7 In Appendix 5, we show that the official debate protocols released by the European Parliament slightly differ from the human transcriptions of speeches, as the final protocols often include corrections made by the EP secretariat or the MEP.

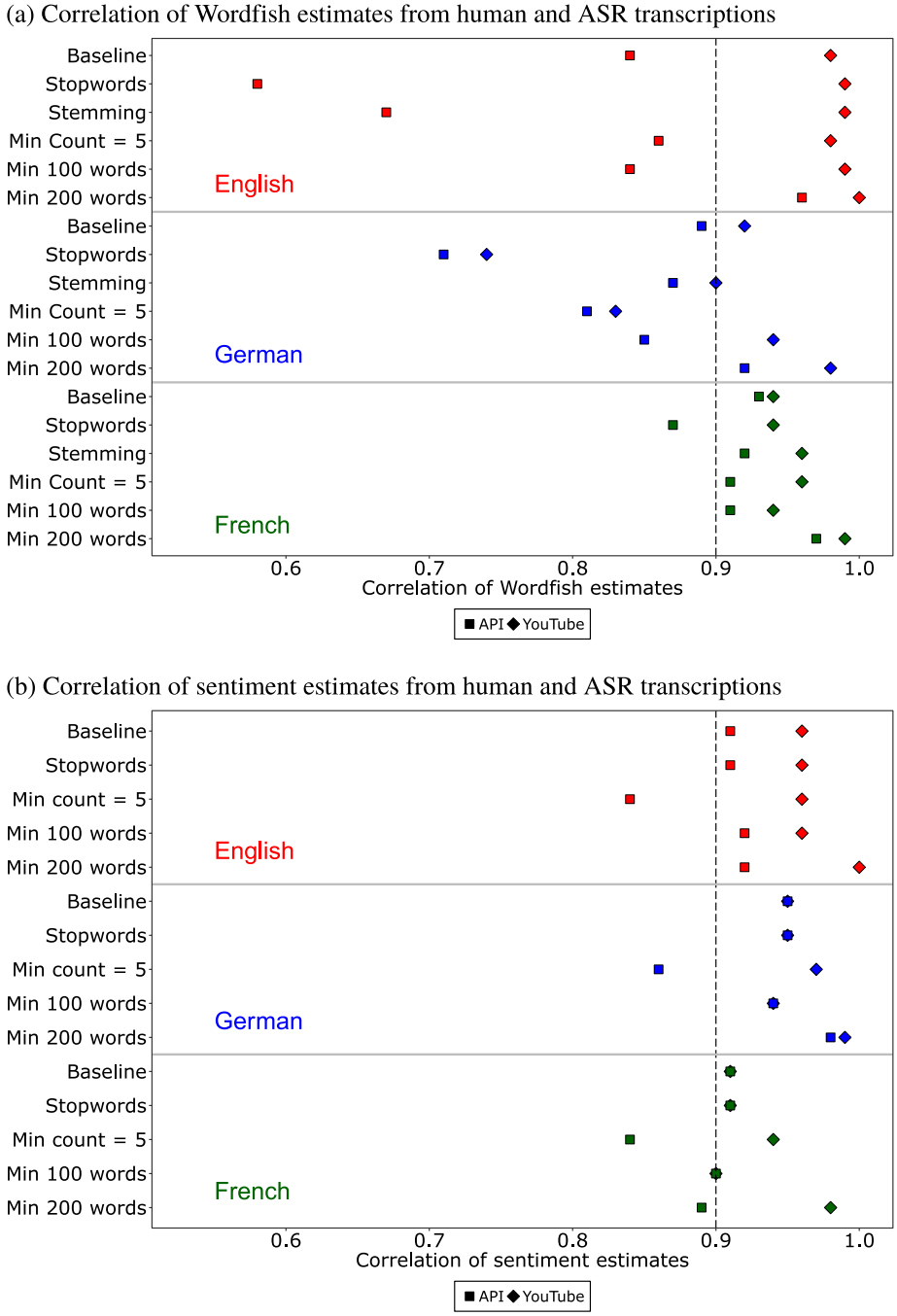

Figure 2. Validity of quantities of interest across languages and ASR types.

3.3 Word Error Rate Simulation (WERSIM)

Since the validity of ASR-based transcriptions in our SOTEU example may not extend to other political text corpora, we suggest the WERSIM procedure that researchers can apply to gauge the sensitivity of their results to the level of transcription error rather than validate the use of ASR for each application.

Assume a quantity of interest Q that is estimated via models involving political speech obtained by ASR. Q may be dictionary counts or position estimates for specific actors from a text scaling model but also some quantity (e.g., a coefficient) from a higher-level analysis model in which estimates from quantitative text analysis are input as data. For instance, the text estimate may

appear as a dependent variable in a regression. We illustrate such cases below, where Wordfish estimates are dependent variables in the final analysis model. In order to gauge the sensitivity of Q to transcription errors, we suggest to simulate how Q changes when we increase the WERs of the ASR texts. Specifically, our procedure consists of the following generic steps:

(1) Simulate M corpora from the original corpus adding transcription errors amounting to an additional WER w, so that the error-inflated corpora have a WER of the original corpus plus w.
(2) Potentially repeat step 1 for different w (e.g., 0.05, 0.10, 0.15, 0.20). Essentially, this adds different amounts of additional word errors, i.e., artificial transcription errors.
(3) Estimate Q for each of the simulated corpora and evaluate the robustness of findings to transcription error at different levels of w (e.g., on the basis of the distribution or the means of Q).

Step 1 poses the problem of how to add transcription errors to the corpus. While in theory the goal is to add errors in a way identical to the error-generating process of the ASR system used, this will often be impossible due to the fact that ASR systems learn independently in real time. Rather than constantly attempting to reverse-engineer the current implementation of the ASR system, we suggest to approximate the types of errors that ASR systems make and examine their implications. As a baseline, for deletions D, we simply randomly draw a unique token from the corpus and delete it from a randomly chosen text it occurs in, and repeat this until we have reached N_D number of deletions needed. In turn, for insertions I, we randomly select a token from the corpus and insert it in a text, repeating this N_I times. Last, we create N_S substitutions S by randomly selecting a unique token and replacing it with the token from the corpus that has the smallest Levenshtein distance (Levenshtein 1966) to the selected token, which measures the similarity between two strings on the basis of single-character differences, e.g., "bad" may be replaced by "bat".

In step 2, to increase the WER by w, we need to insert E errors, with $E = w * N = N_D + N_I + N_S$. In our software implementation (see below), we allow researchers to freely determine the ratio of deletions, insertions and substitutions. We recommend to obtain human transcriptions of a random sample of the corpus, calculate the components of the WER by comparing the human- and the ASR-transcribed corpora, and set the ratios to their empirical estimates. Step 3 allows researchers to evaluate to what extent specific quantities of interest change as a result of increased WERs.

Note that this procedure can be used even if the actual WER of the ASR transcriptions is unknown. The simulation procedure will always add error to what has been returned from the ASR system and allows researchers to examine whether aggregate results change as a result of error and which documents are particularly affected.[8] We provide the R package "wersim" that implements this procedure in various ways for any kind of text model and corpus.[9]

8 Note that while this procedure appears to be similar to a simulation-extrapolation (SIMEX) procedure (Cook and Stefanski 1994), there are important differences. SIMEX corrects for simple additive measurement error of an independent variable in a regression framework. The assumption is that the measurement error of an independent variable has a mean of zero and a fixed variance. Thus, SIMEX can be used to correct for the attenuation effect of measurement error of an independent variable by first simulating additional measurement error (i.e., a greater variance) and then extrapolating back to no measurement error using (mostly) a simple quadratic fit as a conservative estimate (e.g., Benoit, Laver, and Mikhaylov 2009). However, in the case of ASR, the relationship between the WER and Q is more complex: additional word errors may only affect certain documents in some situations and for some quantities of interest. Thus, the additional error will oftentimes be unevenly distributed across documents, leading to complex relationships between WER and Q. Because Q can be much more diverse than just an independent variable with measurement error in a regression framework and since the central assumption of SIMEX of random measurement error does not necessarily hold, we thus do not implement an extrapolation step in WERSIM. Instead, our tool can be used by researchers to explore the robustness of results through simulation for a multitude of research designs.

9 The package is available at https://github.com/jenswaeckerle/wersim.

Table 1. Examining word error rate simulations for the SOTEU debate.

	Dependent variable:	
	IQR for additional 0.2 WER	
	Wordfish model	***Sentiment model***
	[1]	**[2]**
Wordfish absolute deviation from mean	−0.023	
	[−0.168; 0.122]	
Sentiment absolute deviation from mean		0.126
		[0.037; 0.215]
Text length by 1000 words	−0.015	−0.044
	[−0.112; 0.082]	[−0.085; −0.003]
Constant	0.457	0.224
	[0.327; 0.587]	[0.165; 0.282]
Observations	58	58
R^2	0.006	0.179
Adjusted R^2	−0.030	0.150

Note: 95% confidence intervals in parentheses.

3.4 Applying WERSIM: Simulating ASR Error in the SOTEU Debate

Next, we demonstrate how the WERSIM procedure can be used to assess the sensitivity of quantities of interest from the SOTEU debate to transcription errors. Specifically, we simulate $M = 200$ corpora with an additional WER of 0.2 on top of the actual one of our English YouTube corpus.[10] The choice of $w = 0.2$ allows us to assess sensitivity to high levels of transcription errors. We obtain the Wordfish position and sentiment estimates for each speaker for all 200 simulated corpora. We operationalize sensitivity as the interquartile range (IQR) of the distribution of Wordfish or sentiment estimates across simulations.

Which quantities of interest are particularly sensitive to transcription errors? We conjecture that position or sentiment estimates based on short texts should be more sensitive to transcription errors. We are also interested in whether particular values of the quantity of interest, e.g., documents with more extreme positions or sentiment, render it more sensitive to transcription error. In Table 1 we therefore regress the IQR of the Wordfish position and sentiment estimates on the text length and the absolute deviation of the Wordfish position or sentiment from the respective mean. We find no document-level differences in the sensitivity of Wordfish estimates. Neither text length nor the extremity of positions is related to the sensitivity of estimates to transcription errors. In contrast, we find that speakers with more extreme sentiment (either positive or negative) have a higher sensitivity of their sentiment estimates to transcription errors. Longer speeches, on the other hand, lower the sensitivity, as the speech is more likely to contain many occurrences of dictionary terms, meaning that the estimate is sufficiently robust to additional transcription errors.

We contend that such analyses can help researchers when interpreting their results, and they can be performed on any corpus with any text model. For instance, researchers interpreting a sentiment model from the SOTEU debates would be aware now that sentiment estimates from short speeches and those with more extreme values are more sensitive to transcription errors and that they therefore should be more careful, e.g., when assessing rank orders between texts based on ASR transcriptions. In Figure A7 in the Appendix, we show that sentiment estimates

10 We add errors according to the actual distribution of insertions, deletions, and substitutions based on a comparison of the ASR corpus with the corpus of human transcriptions.

from a short speech given by MEP Sophie Auconie vary wildly after introducing additional word errors. The speech is estimated to be close to José Manuel Barroso (a very positive speaker) in some simulations but close to Nigel Farage (a very negative speaker) in others. Hence, statements about the placement of this short text are highly prone to transcription errors. In turn, Barroso's and Farage's longer speeches have much more stable sentiment estimates even after introducing additional word errors.

3.5 Summary

In sum, our results on the SOTEU corpus demonstrate that quantities of interest retrieved from popular quantitative text models using ASR transcriptions converge very strongly with those retrieved using human transcriptions. Typically, correlations of quantities of interest are around 0.95 to 0.99 for English when using YouTube. Moreover, text length improves ASR estimates, and for most political science applications, the texts on which quantities of interest are based are significantly longer than our rather short texts and/or several estimates for (shorter) texts are aggregated into a single variable in the analysis model, which should further reduce measurement error from ASR. Importantly, scholars can use the WERSIM procedure to assess the sensitivity of their quantities of interest to transcription errors. We now proceed to present two novel political text corpora generated via ASR to gain insights into (a) electoral campaigns and (b) budget negotiations in the Council of the European Union.

4 Application: Campaign Dynamics and Party Competition

Campaigns are a central feature of electoral democracy, but their dynamic effects so far remain a largely neglected feature in work on party competition. The standard view suggests that parties adopt policy positions, write manifestos and publish election pledges prior to the election, and that these messages reach voters through party communication (e.g., Fernandez-Vazquez 2014; Thomson *et al.* 2017). There is a vast literature in comparative politics on how parties or candidates adjust their positions in response to competitors in manifestos (e.g., Meguid 2005, 2008; Abou-Chadi 2016). This literature focuses on shifts in positions in the long run (e.g., between elections), but it remains an open question whether parties and candidates perform small position shifts also during the campaign in the short run. For instance, politicians may benefit from marginally adjusting their position during a campaign, virtually on a daily basis, in response to contextual conditions. One prominent case from the world of campaign communications is television debates between lead candidates. On such occasions, politicians may react to characteristics of the competitors they are facing. In principle, the adversarial and accommodative strategies that parties have been found to use in the long run (Meguid 2005) could have micro, short-term counterparts during the campaign, such as when leaders adapt the ideological rhetoric of their speech to the opponents they face in a debate. This phenomenon becomes particularly important as established mainstream parties in Europe are increasingly competing against new populist parties. A conservative politician may try to be accommodative by shifting her rhetoric to the right when debating a right-wing populist, while a socialist politician might employ the opposite strategy and become more adversarial when debating the same right-wing populist.

Up to now it was either expensive or impossible to test such propositions about party competition in the campaign context. To be able to test such expectations we need to rely on a sufficient number of critical campaign events such as television debates or major campaign speeches, which oftentimes are not transcribed and therefore not ready for quantitative analysis. We demonstrate the potential of ASR in this area using the 2017 general election in Austria as a case to study the dynamic positioning of parties during election campaigns. This case is perfectly suited to study campaign dynamics given that Austria's current party system does not only feature a right-wing populist party (the Freedom Party) but also two liberal/green counterparts at the

opposite end of the political spectrum ("NEOS - The New Austria" and the Green Party) that all compete with the "classic" social and Christian democratic mainstream parties. Hence, Austria is emblematic of recent changes in many European party systems (e.g., new emerging niche parties, struggling mainstream parties). In addition, television debates between parties' lead candidates are a central element of the election campaign in Austria, with no less than 23 debates of one to two hours length streamed by different Austrian channels (ORF2, PULS4, and ATV) in the 30 days before the October 2017 election.

Out of these 23 debates, 20 were duels between two lead candidates from different parties, and the remaining three featured all candidates. In general, these debates are structured by moderators posing questions to candidates according to some thematic blocks. However, candidates often delve into topics unrelated to the questions and regularly interrupt each other. We obtain ASR transcriptions of all debates using YouTube as well as candidates' ideological positions from a single Wordfish model using all pooled speech interventions by a candidate in a debate as a document. In preprocessing, we implement the same baseline specifications as above (i.e., lowercasing, number and punctuation removal, minimum feature frequency of three). In Figure 3 we plot the estimated position of each candidate (θ_i), Christian Kern (Social Democrats), Sebastian Kurz (New People's Party), Ulrike Lunacek (Green Party), Hans-Christian Strache (Freedom Party), and Matthias Strolz (NEOS - The New Austria), over the last month of the campaign using a LOESS regression.

The figure reveals that text scaling has a high face validity and accurately recovers the general left–right positions of the Austrian parties in 2017 placing the right-wing populist Freedom Party most to the right, followed by the Christian democratic New People's Party. The Social Democrats occupy the center, and the Green Party and NEOS are placed to the left.[11] In other words, as we would expect, word usage in television debates by lead candidates follows a partisan logic.

While most of the variation in positions is nested *between* candidates, there is obvious residual variation *within* candidates over time. This variation is interesting from a dynamic view on the campaign. Do candidates move their position when they face a particular opponent during a television debate? We investigate this using the Wordfish positions as a dependent variable in linear regression models with fixed effects for candidates (i.e., parties), removing all between-candidate variation. In order to see whether candidates move toward *or* away from their debating partners when communicating to voters on television, we operationalize the debating partners' positions on the basis of expert placements of their parties in the Chapel Hill Expert Survey 2014 (Polk *et al.* 2017). If candidates faced more than one debating partner, we take the arithmetic mean of the debating partners' party positions. The results are shown in Table 2. In Model 1 we test the effect of the debating partners' position measured on the general left–right scale, and in Model 2 we use the positions on the GAL–TAN scale as well as, in Model 3, the positions on an immigration policy scale as potential alternatives. To account for clustering of observations within candidates, we report standard errors from a nonparametric bootstrap on the level of candidates. The results reveal that if candidates face debating partners who are more right/TAN/immigration-restrictive (higher values on the expert ratings), they also position themselves more to the right in their ideological rhetoric during the television debate (higher θ_i). This result is consistent with accommodative strategies, i.e., efforts by the candidate to assimilate her position toward the opponent(s) that confront her in the debate.

Regrettably, our limited sample of 52 observations does not allow us to differentiate this effect further. Hence, we are not able here to test hypotheses about differential use of strategies by different party types (e.g., Meguid 2005; Abou-Chadi 2016). Clearly, our effect could be driven by

11 The placement of the NEOS on the left rather than in the center(-left) suggests that the Wordfish model rather captures a right-authoritarian versus left-libertarian dimension than a purely economic left–right dimension, on which the NEOS are more centrist.

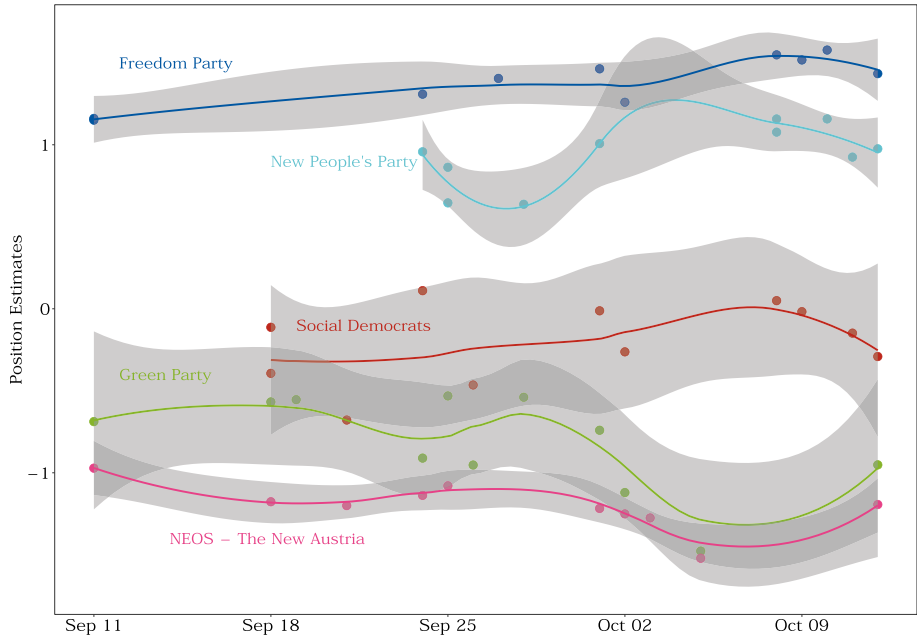

Figure 3. Candidates' positions during the 2017 Austrian Election Campaign. *Note*: Wordfish position estimates (θ_i). Solid lines are LOESS regression lines (span = 0.75). Shaded areas are 95% confidence intervals.

Table 2. Explaining campaign positions of party leaders (Austria 2017).

	Model 1	Model 2	Model 3
Left–right position of debating partners	0.05 [0.03; 0.06]		
Migration policy position of debating partners		0.05 [0.04; 0.07]	
GAL–TAN position of debating partners			0.06 [0.04; 0.08]
Fixed effects	Candidates	Candidates	Candidates
R^2 (within)	0.13	0.31	0.43
N	52	52	52

Note: OLS estimates; 95% confidence intervals in parentheses based on nonparametric bootstrap resampling on the level of candidates (2000 samples).

some combinations of debating partners (e.g., mainstream versus populist right-wing candidate), but on average we see *accommodative* strategies to have prevailed during the 2017 election campaign. Debating the most right-leaning candidate from the Freedom Party compared to the left-leaning candidate from the Greens shifts a candidate's position estimate by about one quarter of a standard deviation to the right.[12]

To assess the robustness of our finding to transcription errors, we apply the WERSIM procedure. Our key quantity of interest is the coefficient estimate for the left–right position of the debating partners reported in Model 1 in Table 2. Hence, we simulate 50 corpora each with 0.05, 0.1, 0.15 and 0.2 additional WER, and re-run the entire analysis including Wordfish and the regression model.

12 We provide several robustness checks of our results in Tables A3 to A5 in the Appendix.

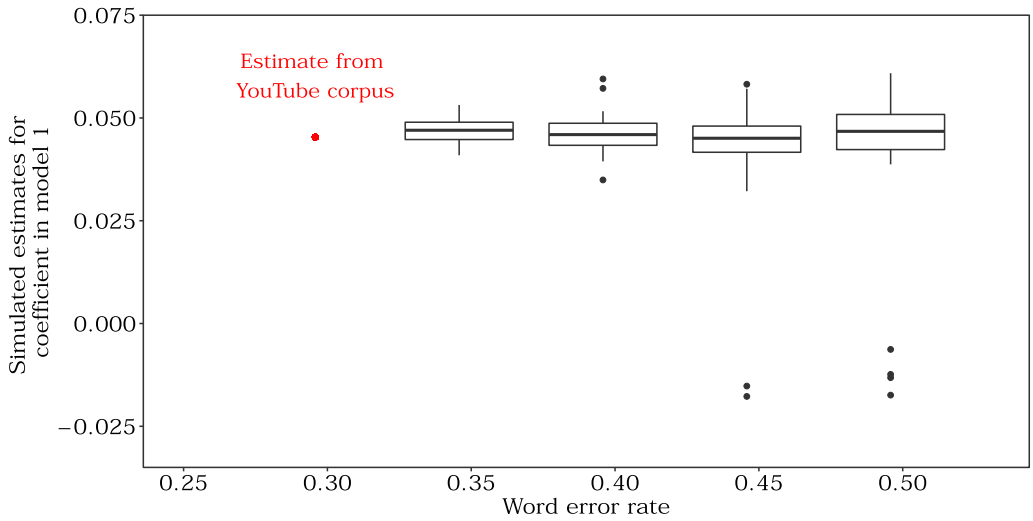

Figure 4. Simulated coefficient estimates for Austria model using WERSIM. *Note*: Box plots show the distribution of estimated coefficients on the left–right position of debating partners from regressions (specified as in Model 1, Table 2) based on each simulated corpus, at additional 0.05, 0.10, 0.15 and 0.20 WER.

Figure 4 displays the distributions of the estimated coefficients at each WER. We find close to no change in the median coefficient in response to additional word errors, and even when adding 0.2 WER, the vast majority of coefficients is positive, which suggests that our finding is robust.

In sum, our analysis shows that an automatically transcribed campaign speech corpus can yield novel insights into campaign dynamics related to partisan position-taking. In particular, we provide some of the first quantitative evidence for dynamic party positions during campaigns when official party positions have already been adopted. We therefore believe that in the future the analysis of such campaign speech corpora can yield new insights for the study of party competition and political communication. Future work could use corpora covering several countries and elections to uncover heterogeneity in parties' strategies.

5 Application: Budget Negotiations in the Council of the EU

Contrary to most national parliamentary chambers, one of the legislative bodies in the European Union does not provide any official verbatim reports of its sessions. The Council of the European Union (henceforth, "the Council"), the EU's main legislative body, represents national governments, with national ministers meeting and negotiating—together with the EP—over EU legislation. While European treaty changes nowadays oblige the Council to make public video footage of all meetings between national ministers that concern legislative deliberations and questions of strategic relevance (European Council 2006), speeches in these meetings are not transcribed.

We demonstrate the opportunities to study such a legislature with ASR transcriptions using the Council's deliberations about the EU's multiannual financial framework (MFF) 2014–2020 as an example. This framework constitutes one of the most important recurring decisions in the EU, as it lays out the long-term budget for a period of seven years defining the level of spending in different areas and programs as well as budget contributions by the member states. Hence, deliberations over this framework can be compared to sessions of national parliaments on annual budgets or long-term budget plans, with the difference that the Council needs to decide *unanimously* on the final agreement. We are interested in whether we can predict the major conflict line on the

basis of the ministers' speeches and discover how the rhetoric in the Council unfolded prior and post agreement.

Budget negotiations in the EU have received broad interest in political science (e.g., Dür and Mateo 2010; Schneider 2011, 2013; Stenbæk and Jensen 2016), but existing studies had to rely on secondary sources, including interviews or budget allocation figures. In turn, using ASR we can analyze speech data from the actual negotiations in the Council. We obtain public videos of all debates in the Council's "General Affairs" configuration on the 2014–2020 MFF.[13] In total, the Council held 17 debates on the framework between July 2011 and December 2016.[14] Ten out of these were held before the European Council's general agreement on the framework on February 8, 2013. In the following, we divide negotiations into a *negotiation phase* (before the general agreement) and an *implementation phase* (after the general agreement). While the negotiation phase focuses on national governments trying to reach a deal, the implementation phase is marked by negotiations with the co-legislator, the EP, and adjustments during the next years. We obtain ASR transcriptions of all deliberations using YouTube.

We are interested in estimating the positions of national governments on the MFF from their ministers' speeches. Regarding face-validating these speeches as a data source, we would expect to see actor alignments that delineate budget contributors from recipient states, i.e., an economic redistribution cleavage. In their study of the 2014–2020 MFF negotiations, Stenbæk and Jensen (2016) find divisions between three groupings of member states: the "friends of better spending" made up of Western European contributors, the "friends of cohesion" composed of Eastern and Southern European recipients, and a group of "friends of agriculture" encompassing member states with mixed budgetary status that were interested in keeping a high agricultural budget. Is the divide between the "friends of better spending" and the "friends of cohesion" (as well as potentially, the "friends of agriculture") also visible in the speech data? Our data coverage of five years and two phases (negotiation, implementation) also allows us to investigate this question dynamically over time. The "friends of better spending" have widely been seen as the "winners" of the 2014–2020 MFF negotiations, since the EU budget has shrunk for the first time in its history (see, e.g., Stenbæk and Jensen 2016). Has this outcome of the negotiation phase subsequently deepened the divide between the two groupings—given that the implementation phase with the inter-institutional agreement and later corrections offered new opportunities to adjust the deal in one or the other direction?

To answer these questions, we obtain Wordfish estimates for each debate using our previous baseline preprocessing specifications for the model (see above). We then use these estimates as manifest variables in two Bayesian factor analysis models. The first model is static and specified following Lauderdale and Herzog's (2016, 378) Wordshoal formulation with:

$$\psi_{i,j} \sim N(\alpha_j + \beta_j \theta_i, \tau_i) \tag{2}$$

$$\theta_i \sim N(0, 1) \tag{3}$$

$$\alpha_j, \beta_j \sim N\left(0, \left(\frac{1}{2}\right)^2\right) \tag{4}$$

$$\tau_i \sim Gamma(1, 1) \tag{5}$$

13 Videos are available at video.consilium.europa.eu.

14 The MFF topic was on one further meeting agenda, but no government except for the presidency took the floor. Hence, we excluded this video.

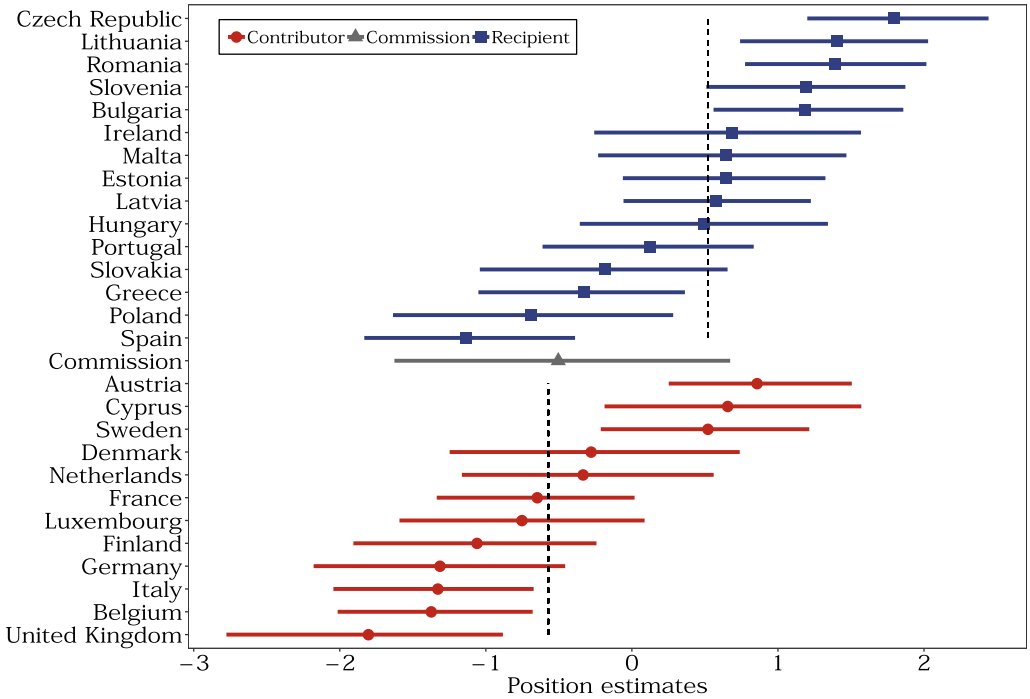

Figure 5. Government position estimates in EU budget negotiations 2011–2016 (Wordshoal). *Note*: Wordshoal position estimates (θ_i). 95% credible intervals as horizontal lines. Dashed vertical lines represent means for contributors and recipients.

where the manifest Wordfish estimates $\psi_{i,j}$ are a linear function of a single latent dimension with member states' position estimates θ_i and debate-specific fixed effects α_j as well as debate loadings β_j, and a normally distributed error (for details, see Lauderdale and Herzog 2016).[15]

Figure 5 shows the estimated governmental positions θ_i. The dimension reveals a clear divide between Western European contributor states (such as the UK, Germany, Belgium or Italy) and Eastern European recipient states (such as the Czech Republic, Lithuania, Romania, or Bulgaria), which broadly overlaps with the distinction between "friends of better spending" and "friends of cohesion".[16] The Commission, which makes the initial proposal for the framework, is placed in the center of the estimated conflict dimension. This suggests it takes a moderating stance, bridging between the two opposing camps. In Appendix 17, we test the robustness of this result using the WERSIM procedure. We introduce additional word error and calculate the difference in means between the positions of contributor and recipient countries in the debates. Figure A19 reveals a high robustness of this result to transcription errors.

We now render the model dynamic in order to see how governmental positions have changed between the negotiation ($t = 1$) and the implementation phase ($t = 2$). For this purpose, we estimate position estimates for each of the two periods $\theta_{i,t}$. We link governments' position estimates over time by centering the standard normal prior for the position estimates in the

15 Note that in order to fix the rotation of the space, we set the mean of the prior on θ_i in the static specification, and on $\theta_{i,1}$ in the dynamic specification, to "−1" for the UK and to "1" for Hungary, deviating from (3) and (6) for these countries. We fit both Bayesian factor models with the JAGS software (Plummer 2003) running one MCMC chain with 1,000,000 iterations and thinning by 100. The JAGS code for both models is in Appendix 12. We assess the convergence of the sampler to its stationary distribution on the basis of the Geweke statistic (see Appendix 13).

16 We take all EU budget figures from www.money-go-round.eu and classify member states as recipients or contributors according to their average annual balance (% of GDP) during the preceding MFF from 2007 to 2013.

Table 3. Increasing polarization in EU MFF negotiations 2011–2016.

	Negotiation ($t = 1$)	**Implementation** ($t = 2$)
Average contributor	−0.52	−0.84
Average recipient	+0.48	+0.72

Note: Cells show the average position estimate for contributor and recipient countries from the dynamic Wordshoal model.

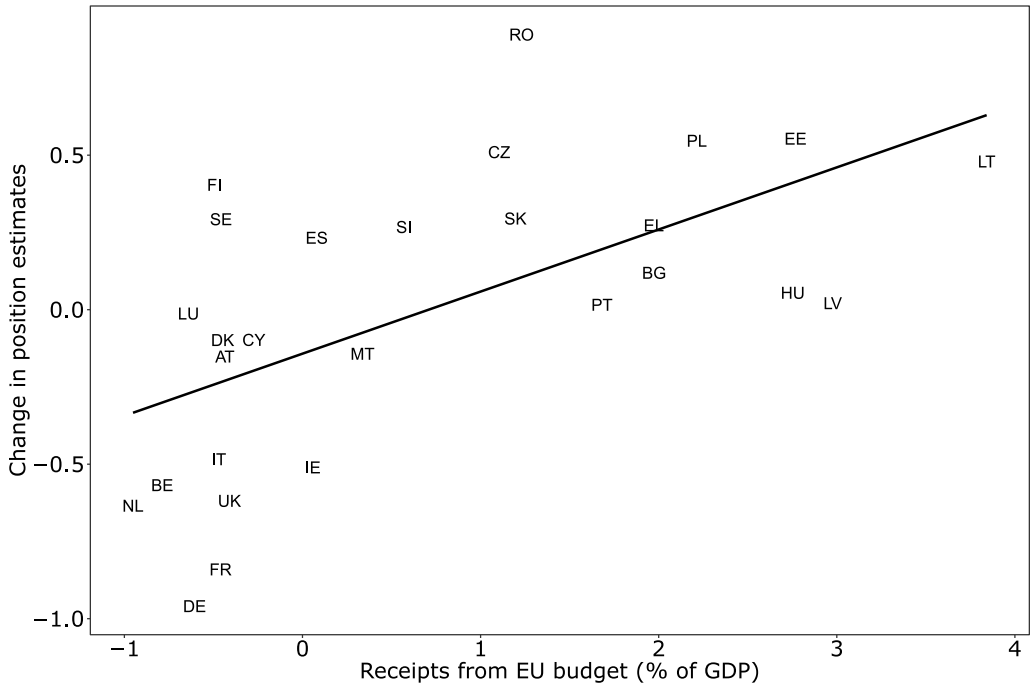

Figure 6. Relationship between change in position and receipts from the EU budget. *Note*: Change in Wordfish position estimates ($\theta_{i,2} - \theta_{i,1}$). See Appendix 11 for country codes.

second period on the position estimates in the first period:

$$\theta_{i,1} \sim N(0, 1) \tag{6}$$

$$\theta_{i,2} \sim N(\theta_{i,1}, 1). \tag{7}$$

This approach of incorporating dynamics was proposed by Martin and Quinn (2002), and subsequently adopted by others (e.g., Schnakenberg and Fariss 2014). Note that we specify a quite large evolution variance (or "smoothing parameter") of "1" in the standard normal prior for the second period, which essentially allows governments to take new positions, largely decoupled from those taken in the negotiation phase. This reflects our aim of investigating possible changes instead of stability, and has little consequences given that we have only two periods and do not compare the amount of movement in different time spans.[17]

How did the positions of governments change between the two periods? In Table 3 we display the average estimated position for contributors versus recipients in the negotiation and the implementation phase. In Figure 6 we also plot the change in position estimates

17 In essence, due to the very weak borrowing between the periods, (5) and (6) are primarily fixing the rotation of the latent space between the periods. We have tested various other specifications of the evolution parameter that all yield substantively the same results on the following analyses.

between the two periods ($\theta_{i,2} - \theta_{i,1}$) against the country's average annual net receipts from the EU budget (in % of GDP) over the period 2007 to 2013 (i.e., the preceding MFF). This reveals that budget contributors and recipients have moved in opposite directions between the negotiation and the implementation phase, resulting in a polarization, i.e., deepening of the division between the "friends of better spending" and the "friends of cohesion". Hence, the redistribution cleavage between member states has intensified in the years following the last long-term budget negotiations.

This application demonstrates the potential for new avenues of studying a political arena which does not publish transcribed protocols but maintains a video library of its negotiations.[18] As we investigated only two time periods, we cannot assess how significant the developing polarization in the EU between countries is compared to common movements in governments' position estimates over time. We leave this question to future research.

6 Practical Considerations

In this section, we provide practical guidance on some of the key considerations involved when researchers decide whether to use ASR transcriptions for a project as well as during the implementation phase of ASR. We focus on when to use ASR and when to use human transcriptions, whether to apply the WERSIM procedure, which ASR system to use, recommendations on implementation, and limitations of the technology for specific research questions.

6.1 Deciding between ASR and Human Transcriptions

Our results provide evidence that there is little reason for using human transcriptions for applications of quantitative text models of the "bag-of-words" variety. In fact, we believe that in terms of the performance of ASR technology the only real decision between human and ASR transcriptions arises if researchers are concerned about grammatical or semantic textual units, i.e., clauses, sentences, or paragraphs. However, with regard to sentences even this limitation may be lifted soon, since the Google API has currently started to offer a beta version including automatic punctuation. A manual solution for obtaining semantic textual units might be to first obtain ASR transcriptions that are then manually corrected by human coders who listen to the audio again. This approach can cut costs (and equivalently human time investment) by about 75% compared to full human transcription from scratch.[19]

6.2 Applying the WERSIM Procedure

Our validation exercise using the SOTEU corpus revealed WERs for transcriptions obtained from YouTube of 0.03 for English, 0.12 for French, and 0.10 for German. This compares to WERs using the Google API of 0.21 for English, 0.26 for French, and 0.21 for German. The corpus of Austrian television debates had a WER of 0.30 and the corpus of debates on the MFF of 0.06 (both using YouTube). This stark variation in WERs in our study suggests that researchers should always explore the sensitivity of their findings to transcription errors. We provide researchers with the "wersim" package in *R* that enables them to do so. Even if the WER of the transcriptions is unknown, researchers can introduce additional error in incremental steps (we use increments of 0.05 WER) and investigate how quantities of interests might change and which texts are most affected by errors. For example, dictionary-based analysis of short texts might be affected by transcription errors as we show in Appendix 4. If researchers have access to human transcriptions

18 To demonstrate the consistency of the results across Google's different ASR technologies, we replicate our analyses using ASR transcriptions of the Council corpus retrieved from the Google API. Comparing the ASR-transcribed corpora to a random sample of human transcriptions yields a WER estimate of 0.06 for YouTube and 0.20 for the API respectively. Despite these varying levels of accuracy, results from the YouTube-based corpus fully replicate with the API-based corpus (see Figures A16 and A17 in the Appendix).

19 We take this figure from the experience of one of our co-authors in another project (for details, see Wratil and Hobolt Forthcoming).

for parts of their corpus, they can calculate the WER to get an idea of the overall error in their transcriptions.

We would like to note that what exactly constitutes robustness will be highly application-specific. In the case of the Austrian election campaign debates, our key interest was to test whether the coefficient on the debating partners' position is positive, since this indicates accommodative strategies. Figure 4 shows that this coefficient is positive in the vast majority of cases, largely independent of the level of transcription errors. In other projects, t-statistics, rank orders, intervals, or sums may be central and researchers should adjust their application of WERSIM to provide the most appropriate robustness check for their core result.

6.3 Research Design and Choice of ASR System

Our results indicate that YouTube provides the most accurate ASR transcriptions, and might therefore be the first choice for researchers. However, the Google API offers some additional features compared to YouTube that might be important in some projects. First, the API provides timestamps for each word, while YouTube only provides timestamps for a couple of words. Second, in contrast to YouTube, the API can be provided with a vector of "speech context", which are words or phrases that help the ASR system in identifying speech. This can be helpful if idiosyncratic vocabulary is especially frequent in a corpus. Third, while YouTube automatically detects the language of a file, the API can be instructed to assume a certain language. This is of importance if the speech file is short or the speakers have a strong accent or dialect, since YouTube then sometimes misinterprets the language. In such cases, the API can often still produce a transcription in the correct language if instructed to do so. These features of the API may be important for some projects and make the API preferable in these cases. Note that the data management and transcription process can be largely automated in the API as well as YouTube, since YouTube captions are also accessible through an API and the "googleLanguageR" (Edmondson 2017) and the "tuber" (Sood 2018) packages allow full access to both interfaces in *R*. Hence, *R* scripts can be used to upload audio-visual files into the cloud or YouTube, send them to the ASR system, and directly build a corpus from the transcriptions.

Sometimes researchers may neither want to rely on the Google API or YouTube due to privacy concerns. One potential example may be sensitive, semistructured interviews with politicians, officials, or experts. In these cases, privacy concerns will usually prohibit uploading the audio files to either the Google cloud or YouTube (not even as "private" content). Nevertheless, for research transparency purposes, transcriptions of interviews would still be helpful. In such cases, an offline ASR system can be used to transcribe the interviews. In case a speaker-dependent offline ASR system must be used to transcribe multiple-speaker files, our experience is that it is most efficient to train the system to an assistant's or the researcher's voice and simply repeat all speech from the audio files to the ASR system. Data may be stored on sites specifically dedicated to storing and sharing interview data in the social sciences (such as the Qualitative Data Repository QDR).

6.4 Awareness of Legal Issues

Before researchers conduct their project they should be fully aware of the legal situation in their country. Are they allowed to create copies of the audio or video content they want to analyze, or is this step potentially in conflict with copyright law? How can a copy be stored and who can have access to it? Are researchers allowed to create transcriptions of the content? Are they allowed to upload the files to either YouTube or the Google cloud or would this potentially be counted as an attempt to make copies available to individuals that should not have access? Can the audio and video files be made available for replication purposes, and to whom? Can the transcriptions be included in replication files? We can just present these as key questions to be answered, since the

specific answers will highly depend on the audio/video content in question as well as the country's copyright laws.

6.5 Implementation of ASR

When implementing ASR researchers will sometimes experience that YouTube (but, to our knowledge, not the API) simply refuses to transcribe the file. While we are not aware of the exact reasons for this, we have found that cutting the file into segments of equal length often solves the issue and each segment is transcribed smoothly. Another option would be to use the API instead of YouTube for this specific file. Despite the occasional benefit of cutting the files, we generally encountered that longer audio files are more often accepted by YouTube than shorter audio files. Hence, as long as no problems occur, researchers have no reason to cut files (e.g., by speaker) before transcription. In our experience, it is best to split the final ASR transcriptions of a longer file (e.g., a whole political debate with several speakers) rather than the audio or video file itself. Researchers should be aware that pure audio files (e.g., from radio interviews) must be converted into a video file if YouTube is used for transcription.

7 Conclusion

Several political debates currently escape systematic scholarly analysis due to the unavailability of transcribed political speech. In this study, we have demonstrated that automatic transcription tools have reached accuracy levels that make them useful for the study of parliamentary debates, campaign speeches, and intergovernmental deliberations. In our applications, using ASR or human transcriptions does not make substantive differences to the results. While we do not claim that quantities of interest generated from ASR transcriptions will always achieve this, we provide researchers with "wersim", an *R* package that allows them to explore the robustness of their findings to transcription errors for any political audio-visual material. Importantly, the most popular ASR systems are very easy to use, demand no specific skills, and virtually any researcher will be able to exploit their services. Only large-scale use of ASR APIs may demand some minimal programming skills.

We presented evidence for two types of applications in which the use of automatic transcriptions can yield novel insights. The first uses a text corpus generated from television debates during an election campaign. This corpus allows us to estimate a time series of candidate positions that other data sources so far have failed to provide. Our analysis of estimated candidate positions yields evidence that candidate positions tend to follow the party manifesto, but that they are dynamic during the campaign. In particular, we found evidence that candidates take accommodative stances. In a second application, we analyzed ministerial speeches in the Council of the EU, a legislative chamber with no official transcriptions of its proceedings. Using long-term budget negotiations as an example, we were able to recover a redistributive conflict dimension between member states. In addition, we could show that polarization levels between member states have actually increased after the political agreement on the budget, thus possibly solidifying the latent, redistributive conflict in the EU.

A limitation of our study is that we have explored a limited number of text models that we have used extensively in our own previous research. However, we provide the tools and code necessary to validate the use of ASR in other applications with other text models, in particular, with the help of the WERSIM procedure. Using ASR technology can yield valid quantities of interest and should open up new research opportunities for comparative and international relations scholars alike.

Supplementary material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2018.62.

References

Abou-Chadi, T. 2016. "Niche Party Success and Mainstream Party Policy Shifts—How Green and Radical Right Parties Differ in their Impact." *British Journal of Political Science* 46(2):417–436.

Accornero, G., and P. Ramos Pinto. 2015. "Mild Mannered? Protest and Mobilisation in Portugal Under Austerity, 2010–2013." *West European Politics* 38(3):491–515.

Baum, M. A. 2005. "Talking the Vote: Why Presidential Candidates Hit the Talk Show Circuit." *American Journal of Political Science* 49(2):213–234.

Baum, M. A., and A. S. Jamison. 2006. "The Oprah Effect: How Soft News Helps Inattentive Citizens Vote Consistently." *Journal of Politics* 68(4):946–959.

Benoit, K., M. Laver, and S. Mikhaylov. 2009. "Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions." *American Journal of Political Science* 53(2):495–513.

Cook, J. R., and L. A. Stefanski. 1994. "Simulation-Extrapolation Estimation in Parametric Measurement Error Models." *Journal of the American Statistical Association* 89(428):1314–1328.

Daku, M., S. Soroka, and L. Young. 2015. "Lexicoder, version 3.0." www.lexicoder.com.

de Vries, E., M. Schoonvelde, and G. Schumacher. 2018. "No Longer Lost in Translation. Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications." *Political Analysis* 26(4):417–430.

Denny, M. J., and A. Spirling. 2018. "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do About It." *Political Analysis* 26(2):168–189.

Dietrich, B. J., M. Hayes, and D. Z. O'Brien. 2017. "Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech on Women." Working Paper.

Dietrich, B. J., R. D. Enos, and M. Sen. 2018. "Emotional Arousal Predicts Voting on the U.S. Supreme Court." *Political Analysis*, https://doi.org/10.1017/pan.2018.47.

Dolezal, M., S. Hutter, and R. Becker. 2016. "Protesting European Integration: Politicisation from Below? In *Politicising Europe: Integration and Mass Politics*, edited by Swen Hutter, Edgar Grande, and Hanspeter Kriesi, 112–136. Cambridge: Cambridge University Press, chapter 5.

Dresing, T., T. Pehl, and C. Schmieder. 2015. "Manual (on) Transcription. Transcription Conventions, Software Guides and Practical Hints for Qualitative Researchers." 3rd English Edition: Marburg. http://www.audiotranskription.de/english/transcription-practicalguide.htm.

Dür, A., and G. Mateo. 2010. "Bargaining Power and Negotiation Tactics: The Negotiations on the EU's Financial Perspective, 2007–2013." *Journal of Common Market Studies* 48(3):557–578.

Edmondson, M. 2017. "Package 'googleLanguageR'." Technical report.

European Council 2006. "An Overall Policy on Transparency." http://www.consilium.europa.eu/ueDocs/cms_Data/docs/pressData/en/misc/90112.pdf.

Fernandez-Vazquez, P. 2014. "And Yet it Moves: The Effect of Election Platforms on Party Policy Images." *Comparative Political Studies* 47(14):1919–1944.

Gibson, R. K., and I. McAllister. 2011. "Do Online Election Campaigns Win Votes? The 2007 Australian YouTube Election." *Political Communication* 28(2):227–244.

Giugni, M., and M. Grasso. 2015. *Austerity and Protest: Popular Contention in Times of Economic Crisis*. Routledge.

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

Grimmer, J., and B. M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3):267–297.

Haynes, A. A., and B. Pitts. 2009. "Making an Impression: New Media in the 2008 Presidential Nomination Campaigns." *PS—Political Science and Politics* 42(1):53–58.

Herzog, A., and K. Benoit. 2015. "The Most Unkindest Cuts: Speaker Selection and Expressed Government Dissent During Economic Crisis." *The Journal of Politics* 77(4):1157–1175.

Hinton, G., L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. 2012. "Deep Neural Networks for Acoustic Modeling in Speech Recognition." *IEEE Signal Processing Magazine* (November):82–97.

Hochreiter, S., and J. Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9(8):1735–1780.

Këpuska, V., and G. Bohouta. 2017. "Comparing Speech Recognition Systems (Microsoft API, Google API and CMU Sphinx)." *International Journal of Engineering Research and Applications* 7(3):20–24.

Knox, D., and C. Lucas. 2018. "A Dynamic Model of Speech for The Social Sciences." Working Paper.

Kriesi, H., E. Grande, M. Dolezal, M. Helbling, D. Hoglinger, S. Hutter, and B. Wueest. 2012. *Political Conflict in Western Europe*. Cambridge: Cambridge University Press.

Lauderdale, B. E., and A. Herzog. 2016. "Measuring Political Positions from Legislative Speech." *Political Analysis* 24(3):374–394.

Lecun, Y., Y. Bengio, and G. Hinton. 2015. "Deep Learning." *Nature* 521(7553):436–444.

Levenshtein, V. I. 1966. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals." *Soviet Physics Doklady* 10(8):707–710.

Liao, H., E. McDermott, and A. Senior. 2013. "Large Scale Deep Neural Network Acoustic Modeling with Semi-Supervised Training Data for YouTube Video Transcription." In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013—Proceedings*, 368–373.
Martin, A. D., and K. M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for The U.S. Supreme Court, 1953–1999." *Political Analysis* 10(2):134–153.
McKibben, H. E. 2016. "To Link or Not to Link? Agenda Change in International Bargaining." *British Journal of Political Science* 46(2):371–393.
Meeker, M. 2017. "Kleiner Perkins Internet Trends 2017." Technical report. http://www.kpcb.com/internet-trends.
Meguid, B. M. 2005. "Competition Between Unequals: The Role of Mainstream Party Strategy in Niche Party Success." *American Political Science Review* 99(3):347–359.
Meguid, B. M. 2008. *Party Competition between Unequals: Strategies and Electoral Fortunes in Western Europe*. Cambridge: Cambridge University Press.
Plummer, M. 2003. "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling." In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, 20–22.
Polk, J., J. Rovny, R. Bakker, E. Edwards, L. Hooghe, S. Jolly, J. Koedam, F. Kostelka, G. Marks, G. Schumacher, M. Steenbergen, M. Vachudova, and M. Zilovic. 2017. "Explaining the Salience of Anti-Elitism and Reducing Political Corruption for Political Parties in Europe with the 2014 Chapel Hill Expert Survey Data." *Research & Politics* 1–9.
Proksch, S.-O., W. Lowe, J. Wäckerle, and S. Soroka. 2018. "Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Parliamentary Speeches." *Legislative Studies Quarterly*, https://doi.org/10.1111/lsq.12218.
Proksch, S.-O., C. Wratil, and J. Wäckerle. 2018. "Replication Data for: 'Testing the Validity of Automatic Speech Recognition for Political Text Analysis' by Sven-Oliver Proksch, Christopher Wratil and Jens Wäckerle." https://doi.org/10.7910/DVN/PGQY2F, Harvard Dataverse, V1.
Proksch, S. O., and J. B. Slapin. 2012. "Institutional Foundations of Legislative Speech." *American Journal of Political Science* 56(3):520–537.
Proksch, S. O., and J. B. Slapin. 2015. *The Politics of Parliamentary Debate: Parties, Rebels, and Representation*. Cambridge: Cambridge University Press.
Proksch, S.-O., W. Lowe, J. Wäckerle, and S. Soroka. 2018. "Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Parliamentary Speeches." *Legislative Studies Quarterly*, https://doi.org/10.1111/lsq.12218.
Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4):1064–1082.
Schnakenberg, K. E., and C. J. Fariss. 2014. "Dynamic Patterns of Human Rights Practices." *Political Science Research and Methods* 2(1):1–31.
Schneider, C. J. 2011. "Weak States and Institutionalized Bargaining Power in International Organizations." *International Studies Quarterly* 55(2):331–355.
Schneider, C. J. 2013. "Globalizing Electoral Politics: Political Competence and Distributional Bargaining in the European Union." *World Politics* 65(3):452–490.
Slapin, J. B., and S.-O. Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3):705–722.
Soltau, H., H. Liao, and H. Sak. 2017. "Neural Speech Recognizer: Acoustic-to-Word LSTM Model for Large Vocabulary Speech Recognition." In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, August*, 3707–3711.
Sood, G. 2018. "Package 'tubr'." Technical report.
Stenbæk, J., and M. D. Jensen. 2016. "Evading the Joint Decision Trap: The Multiannual Financial Framework 2014–20." *European Political Science Review* 8(4):615–635.
Thomson, R., T. Royed, E. Naurin, J. Artés, R. Costello, L. Ennser-Jedenastik, M. Ferguson, P. Kostadinova, C. Moury, F. Pétry, and K. Praprotnik. 2017. "The Fulfillment of Parties' Election Pledges: A Comparative Study on the Impact of Power Sharing." *American Journal of Political Science* 61(3):527–542.
Wratil, C., and S. B. Hobolt. Forthcoming. "Public Deliberations in The Council of the European Union: Introducing and Validating the DICEU Approach." *European Union Politics* 1–27.
Young, L., and S. Soroka. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts." *Political Communication* 29(2):205–231.
Yu, D., and L. Deng. 2015. *Automatic Speech Recognition*. Signals and Communication Technology London. London: Springer.
Zittel, T. 2015. "Do Candidates Seek Personal Votes on the Internet? Constituency Candidates in the 2009 German Federal Elections." *German Politics* 24(4):435–450.