Testing the Validity of Automatic Speech Recognition for Political Text Analysis Appendix

Political Analysis

Sven-Oliver Proksch
University of Cologne

Christopher Wratil
Harvard University
and University of Cologne

Jens Wäckerle
University of Cologne

November 28, 2018

Contents

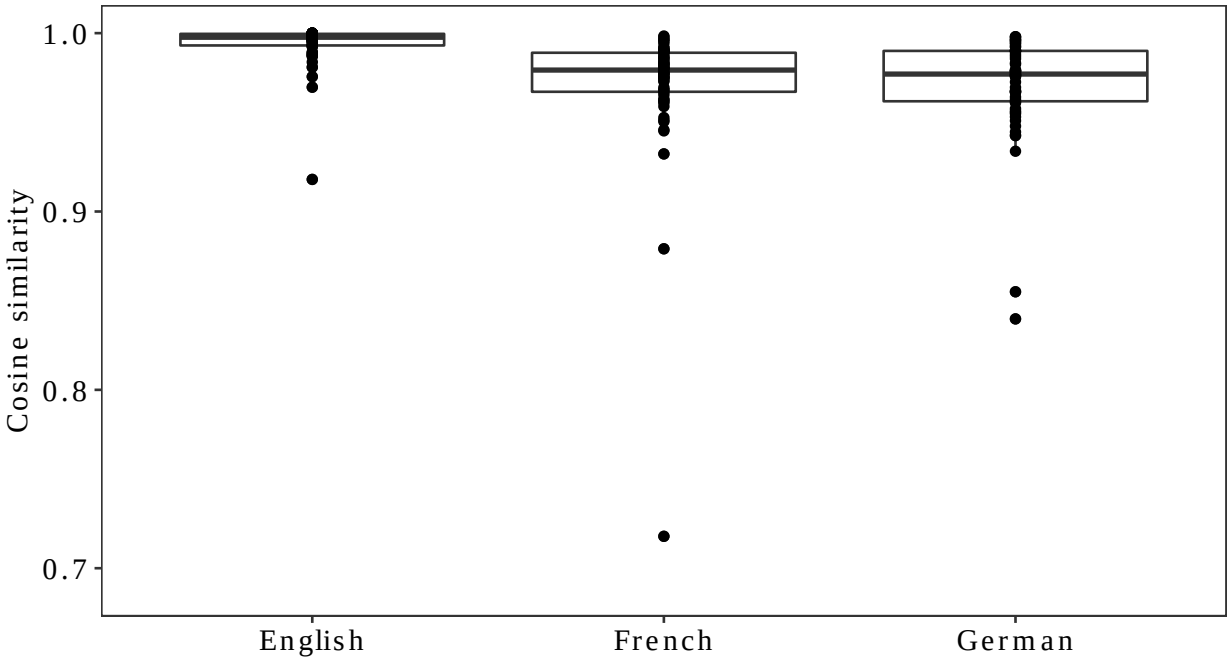
List of Figures

List of Tables

1 Cosine Similarity for State of the Union Speakers

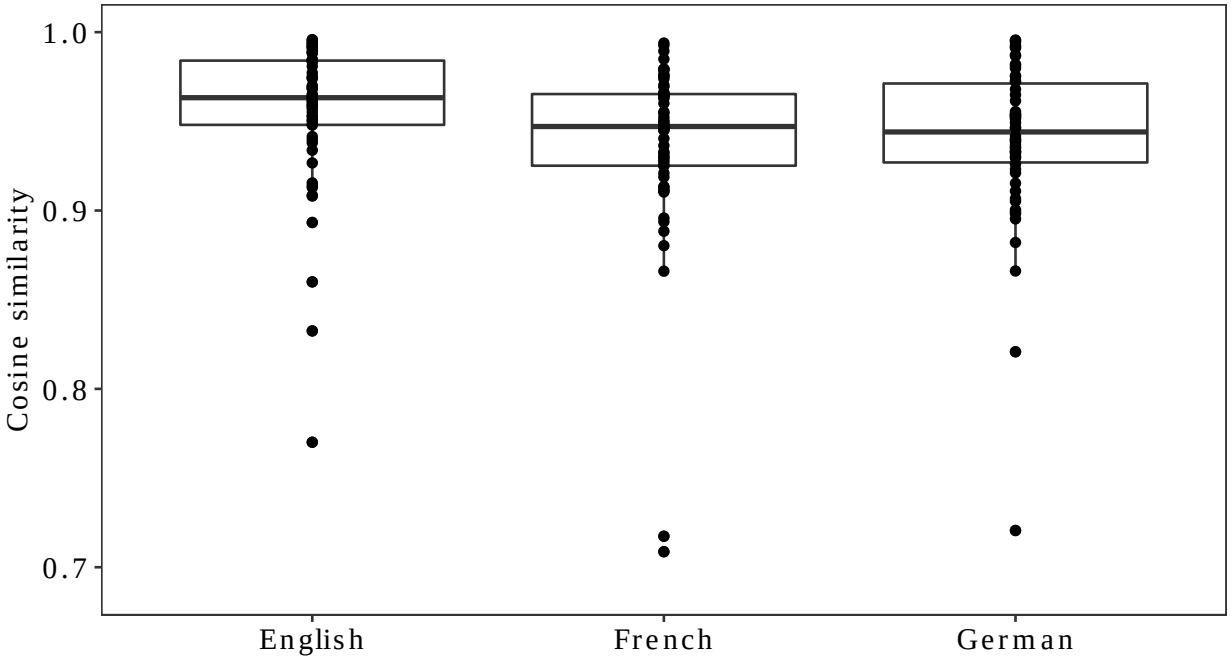
Cosine similarity is calculated as $cos(x, y) = \frac{x*y}{||x||*||y||}$ for two documents x and y. With text vectors the values of this measure can range from 0 to 1, where 1 denotes complete similarity. We calculate similarity seperately for each pair of speeches from the "word-by-word" corpus and the corpus of ASR transcriptions by YouTube (Figure A1) as well as the API (Figure A2). Similarities are generally very high. We checked all outliers and verified that these are very short texts (all $<$ 150 words).

Figure A1: Cosine Similarity for State of the Union Speakers - YouTube



Note: Boxplots plotting the median, 25th and 75th percentiles, whiskers that extend to 1.5 * IQR as well as measurements for individual texts.

Figure A2: Cosine Similarity for State of the Union Speakers - API

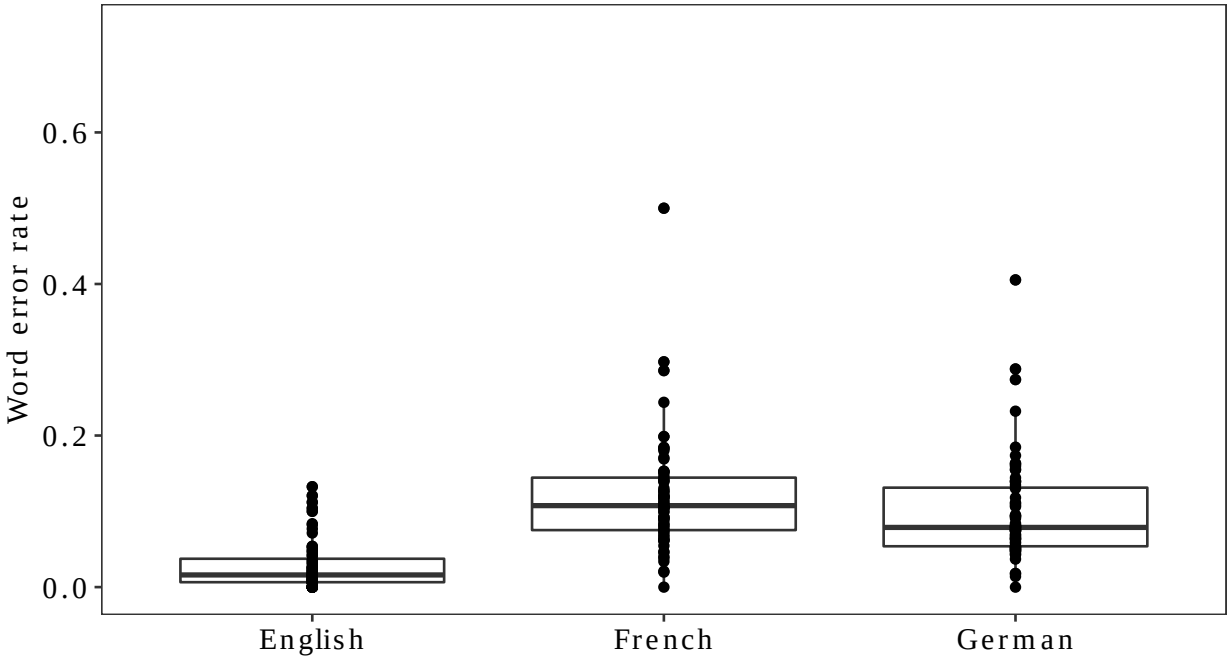


Note: Boxplots plotting the median, 25th and 75th percentiles, whiskers that extend to 1.5 * IQR as well as measurements for individual texts.

2 Word Error Rate for State of the Union Speakers

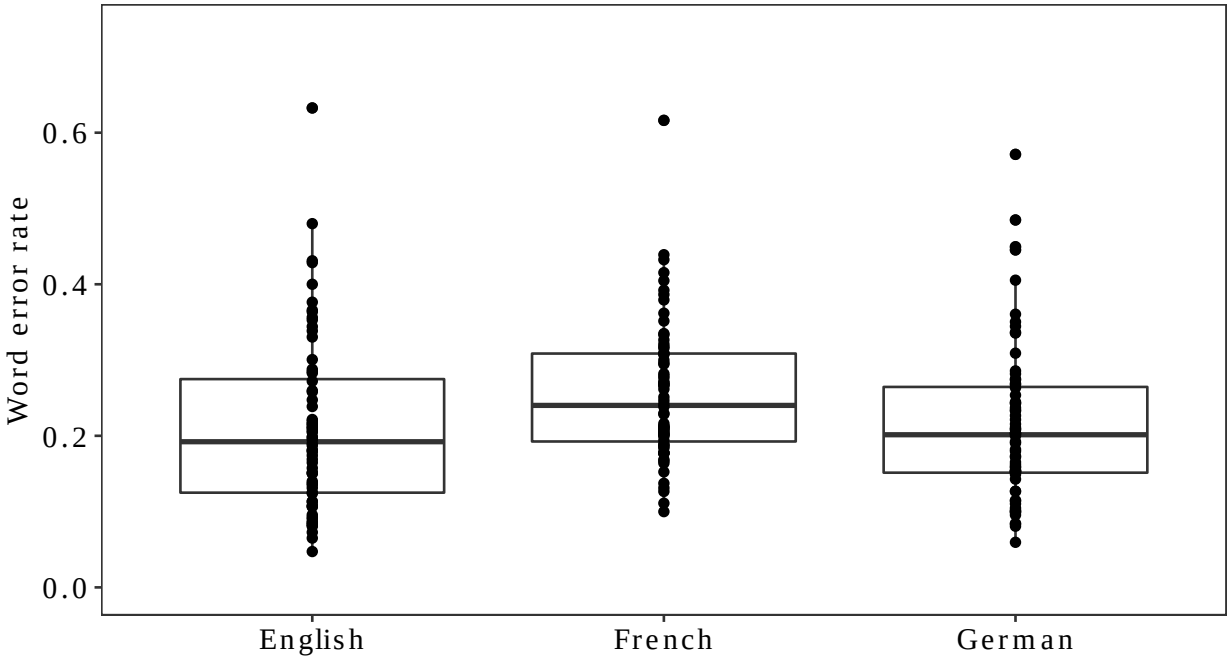
In Figures A3 and A4 we plot the WERs for each speaker in the SOTEU debate (comparing "word-by-word" to ASR transcriptions). We investigated all outliers, which stem mostly from very short texts (< 100 words).

Figure A3: Word Error Rate for State of the Union Speakers - YouTube



Note: Boxplots plotting the median, 25th and 75th percentiles, whiskers that extend to 1.5 * IQR as well as measurements for individual texts.

Figure A4: Word Error Rate for State of the Union Speakers - API



Note: Boxplots plotting the median, 25th and 75th percentiles, whiskers that extend to 1.5 * IQR as well as measurements for individual texts.

3 Wordfish Estimates of Speakers and Parties for the 2011 State of the Union

In order to illustrate the substantive meaning of our SOTEU Wordfish models from Figure 2 in the paper, we plot speaker and party position estimates (pooling all speeches by members of a party) in Figures A5 and A6 using the English language corpus and our baseline pre-processing specifications. This reveals a political space with the European Commission on the one end of the spectrum and the Eurosceptic GUE/NGL and EFD at the other end. The dimension well reflects pro-anti integration as well as mainstream-niche party differences.

Figure A5: Wordfish Estimates of Speakers for the 2011 State of the Union (English Corpus)

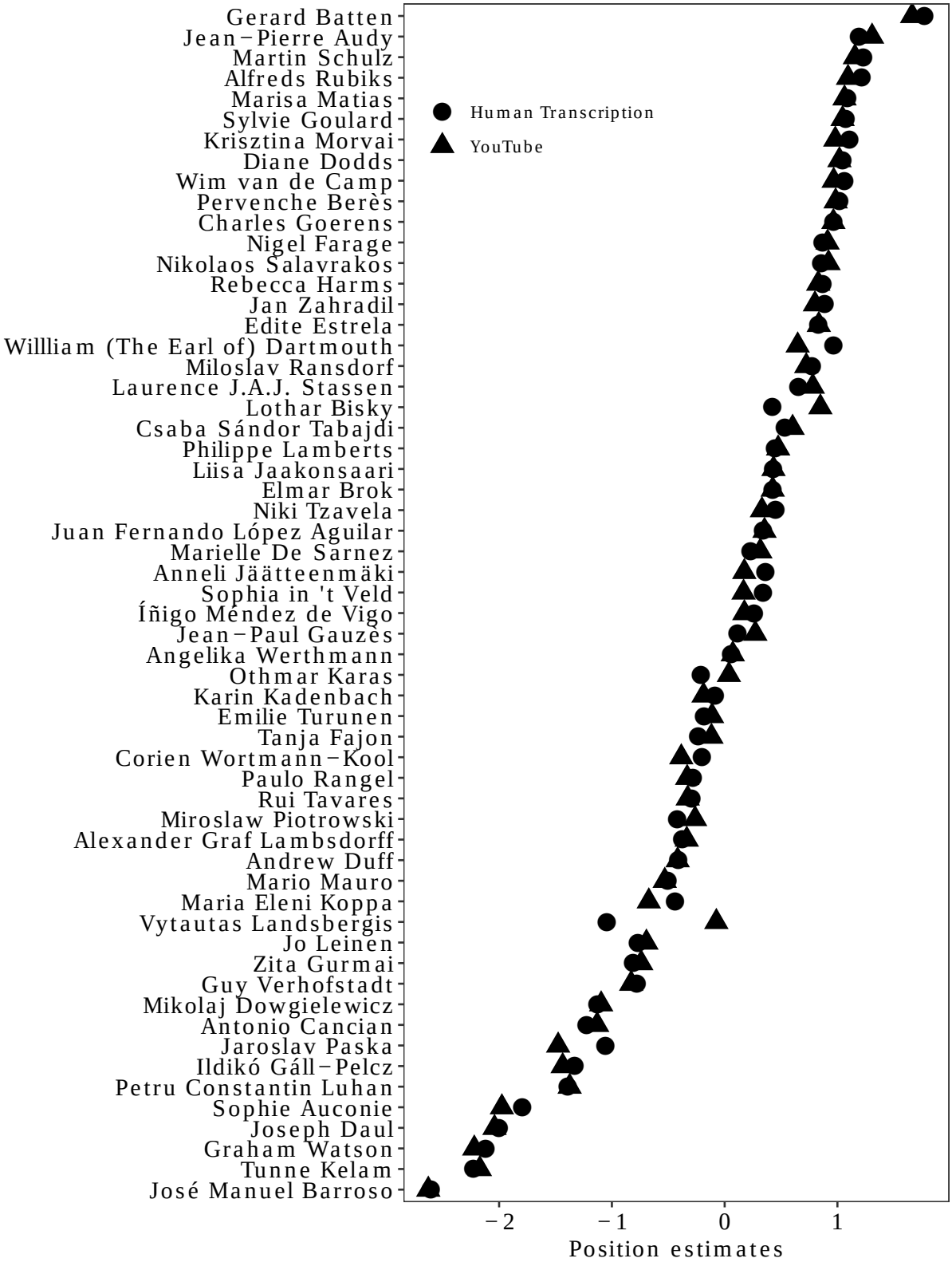
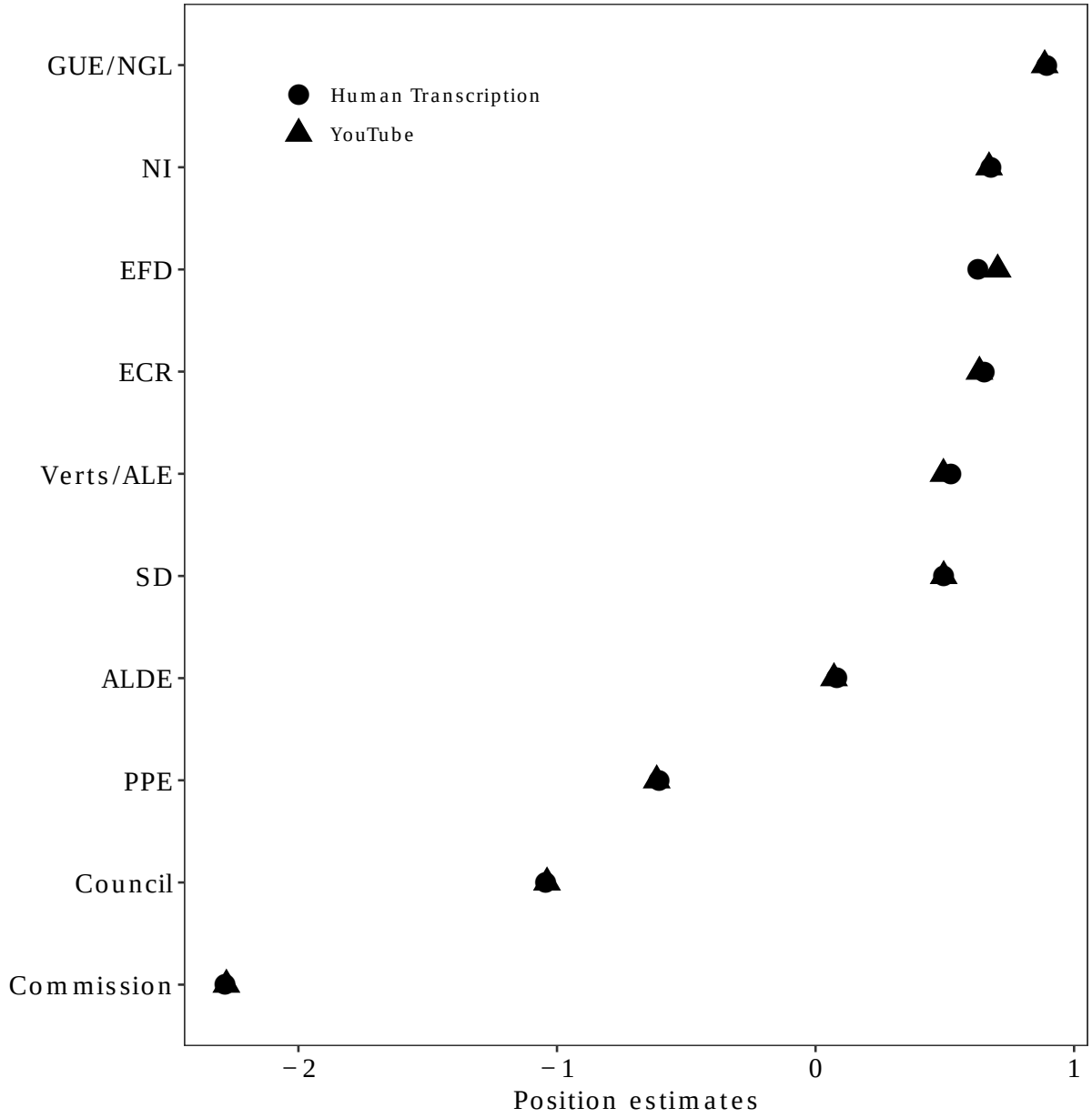


Figure A6: Wordfish Estimates of Parties for the 2011 State of the Union (English Corpus)

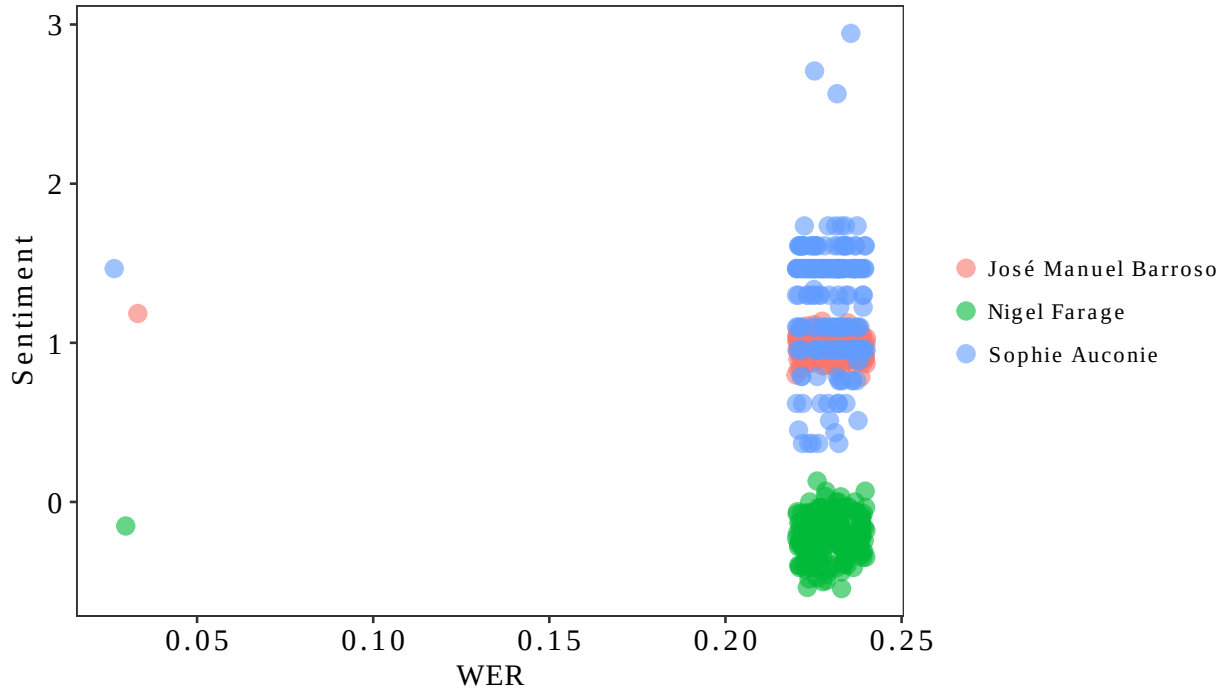


Note: GUE/NGL: European United LeftNordic Green Left; NI: Non-attached members; EFD: Europe of Freedom and Direct Democracy; ECR: European Conservatives and Reformists; Verts/ALE: The Greens/European Free Alliance; SD: Progressive Alliance of Socialists and Democrats; ALDE: Alliance of Liberals and Democrats for Europe; EPP: European People's Party.

4 Simulated Sentiment Estimates from the WERSIM Procedure

Figure A7 shows the sentiment estimates for three speakers (José Manuel Barroso, Nigel Farage and Sophie Auconie) in the YouTube corpus and in 200 corpora that were simulated with a WER of 0.2259 using the WERSIM method. As Barroso has the longest speech, his estimate is fairly stable after introducing transcription error. Auconie, on the other hand, has a very short speech and her estimate varies wildly after introducing error.

Figure A7: Simulated Sentiment Estimates with the WERSIM Procedure

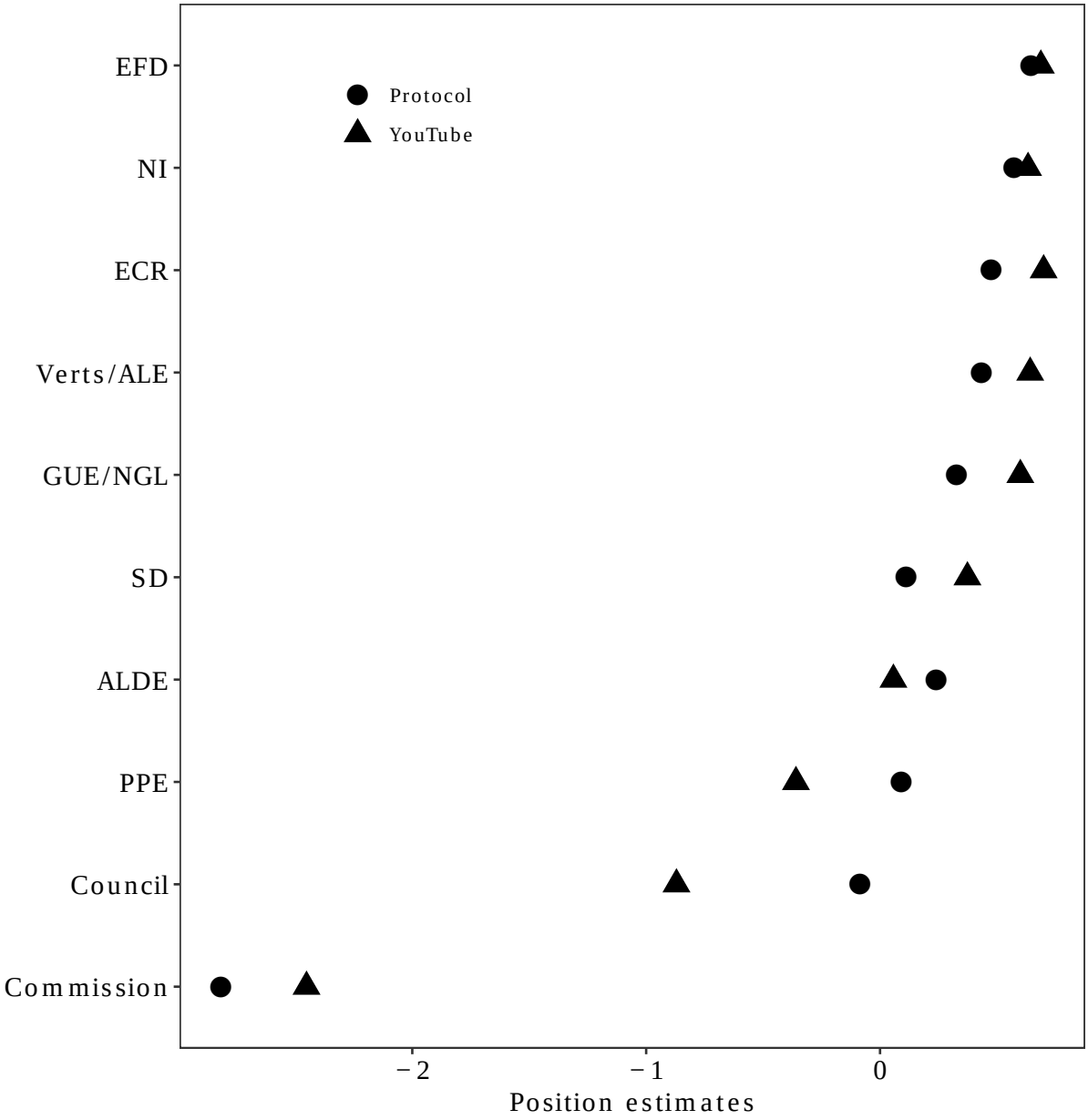


Note: Estimates are shown using the jitter function, the three estimates on the left are exactly at WER of 0.0259 and on the right at WER of 0.2259.

5 Analysis of Official Protocols for SOTEU debate

While we rely on parliamentary debates in the EP for our validation of ASR systems, we note that researchers will use the actual protocols when analyzing parliamentary speech in their projects whenever available. Verbatim protocols of the EP, as well as other parliaments like the UK House of Commons or the German Bundestag, are often not identical to the spoken words in the plenary session, since repetitions and obvious mistakes by speakers are usually corrected by stenographers and members can request post-hoc alterations to the reports. For instance, in the SOTEU corpus the WER of verbatim reports for speeches held in English when using our "word-by-word" human transcriptions as reference texts is 0.27. The verbatim reports are therefore different from human "word-by-word" transcriptions. We compare the substantive results for protocols to ASR transcriptions in Figure A8. They show that ASR almost perfectly recovers the spoken words, while the protocols contain modifications that lead to slightly different position estimates. Nevertheless the placement of political groups correlates highly at 0.93 between YouTube and the protocol corpus.

Figure A8: Wordfish Estimates of Parties for the 2011 State of the Union (English Corpus), with protocol



Note: GUE/NGL: European United LeftNordic Green Left; NI: Non-attached members; EFD: Europe of Freedom and Direct Democracy; ECR: European Conservatives and Reformists; Verts/ALE: The Greens/European Free Alliance; SD: Progressive Alliance of Socialists and Democrats; ALDE: Alliance of Liberals and Democrats for Europe; EPP: European People's Party.

6 Sentiment Estimates of Speakers and Parties for the 2011 State of the Union

We also present the sentiment estimates by speaker and party in Figures A9 and A10. This reveals a very similar order of parties compared to the political space scaled with Wordfish (see above). The European Commission talks most positively and the ECR most negatively.

Figure A9: Sentiment Estimates of Speakers for the 2011 State of the Union (English Corpus)

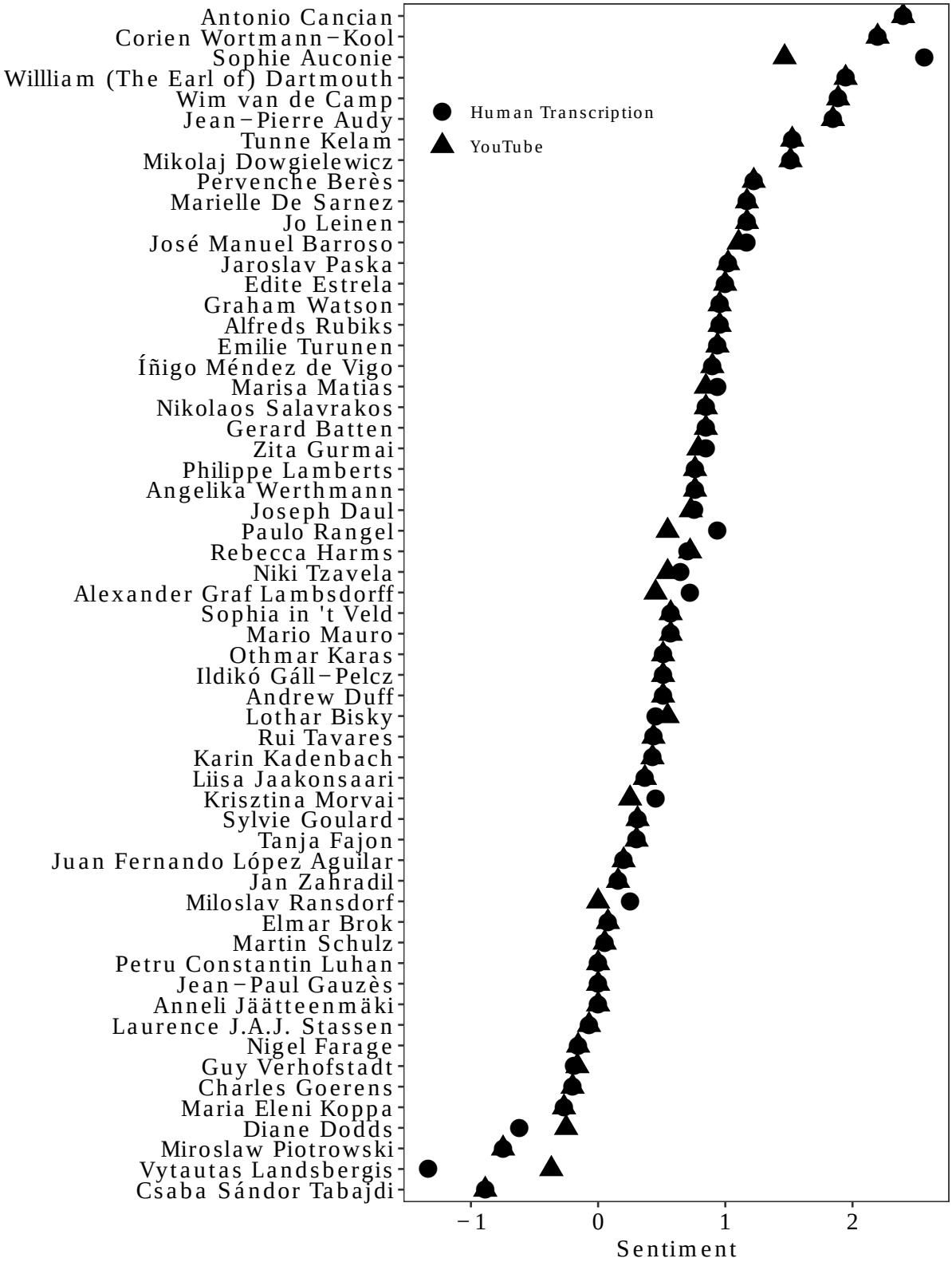
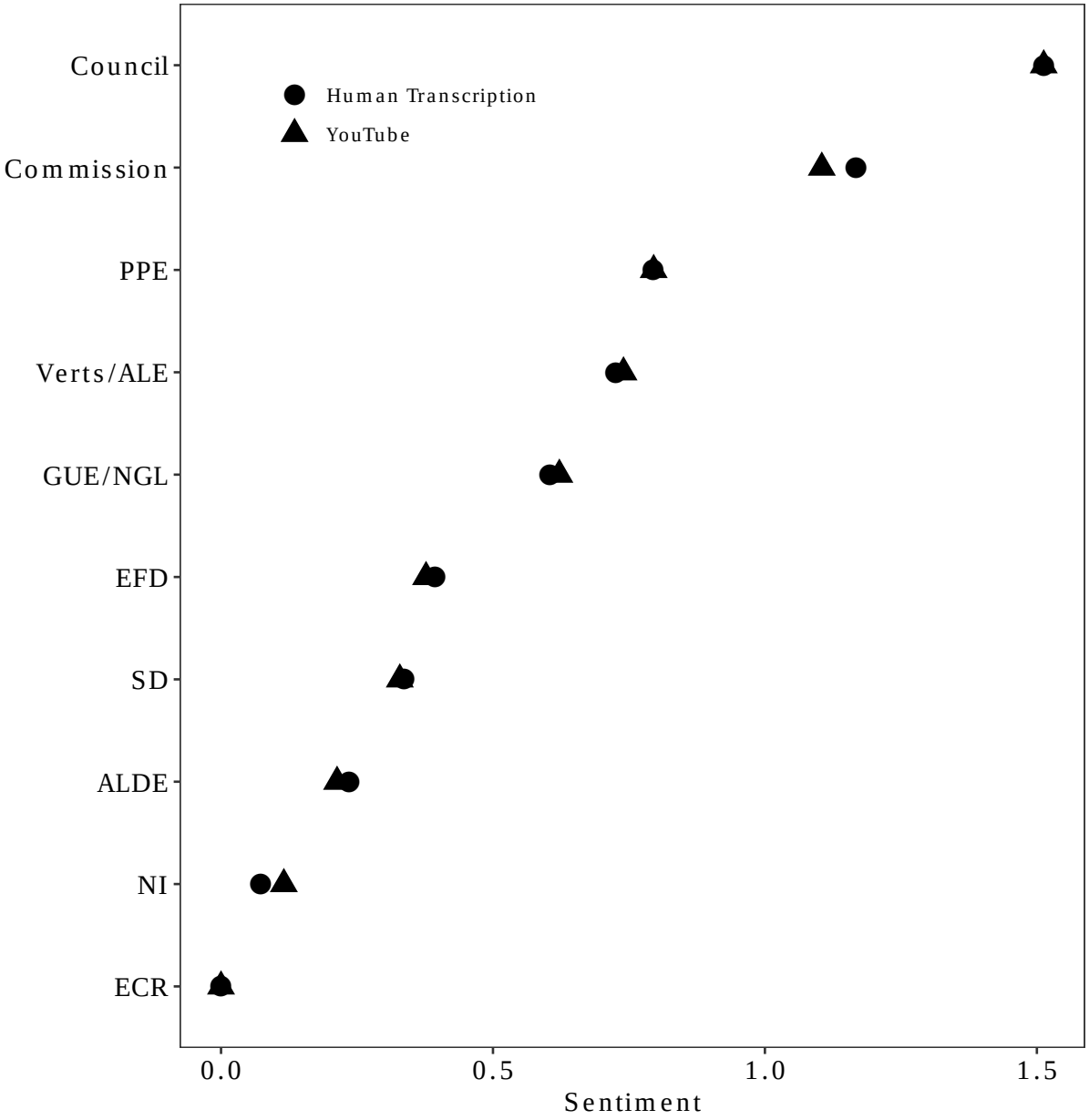


Figure A10: Sentiment Estimates of Parties for the 2011 State of the Union (English Corpus)

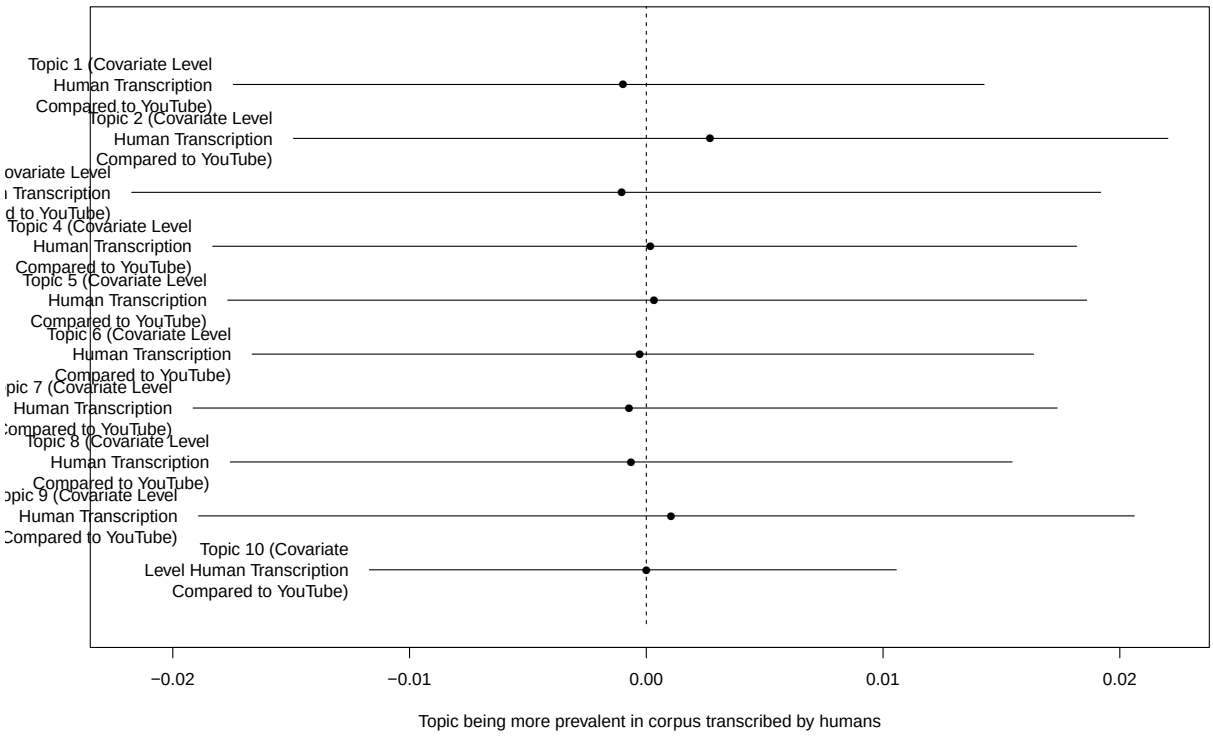


Note: GUE/NGL: European United LeftNordic Green Left; NI: Non-attached members; EFD: Europe of Freedom and Direct Democracy; ECR: European Conservatives and Reformists; Verts/ALE: The Greens/European Free Alliance; SD: Progressive Alliance of Socialists and Democrats; ALDE: Alliance of Liberals and Democrats for Europe; EPP: European People's Party.

7 Structural Topic Models of the State of the European Union Debate

We ran structural topic models (stm) on the State of the European Union corpus to test whether our automated transcriptions produce similar results as the human word-by-word transcriptions for this kind of analysis. As the number of texts included in the corpus is rather small and stm is commonly used for large corpora with many individual texts, we split the corpora in twenty-word increments. This produced about 1000 short texts for each of the two corpora which we then combined and sublemented with a covariate indicating whether the text came from the YouTube-transcribed or human-transcribed corpus. We would want to find no effect of mode of transcription on topic prevalence, so no topic being more or less likely to occur in one of the corpora. This is in fact what we find, as Figure A11 shows.

Figure A11: Effect of Transcription Mode on Topic Prevalence in STM Model of the SOTEU Debate



8 Correlation of Sentiment Estimates with Short Dictionary

In order to assess whether ASR transcriptions are valid for shorter dictionaries than the sentiment dictionaries we use in the paper, we simulate 1,000 *short* sentiment dictionaries per language by 1,000 times randomly drawing 10% of all words from the original dictionaries. The simulated short dictionaries contain 455 words for English, 3120 words for German, and 414 words for French. As in Figure 2 in the paper, we compare sentiment estimates using these short dictionaries for ASR and human "word-by-word" transcriptions. Table A1 shows that the average correlations of the sentiment estimates (as quantity of interest) from shorter dictionaries are generally still high (especially, when using YouTube transcriptions). We also report the values of the 5th and the 95th percentiles of these correlations from our 1,000 simulated dictionaries.

Table A1: Correlation of Sentiment Estimates with Short Dictionary

Original English	Original French	Original German	
0.98			English YouTube
[0.94; > 0.99]			
0.85			English API
[0.74; 0.93]			
	0.92		French YouTube
	[0.86; 0.96]		
	0.81		French API
	[0.69; 0.90]		
		0.92	German YouTube
		[0.86; 0.97]	
		0.87	German API
		[0.78; 0.94]	

Note: Based on 1,000 simulated dictionaries per language. 5th and 95th percentiles of correlation estimates in parentheses.

9 Austrian TV Debate Schedule

Table A2: Austrian TV Debate Schedule

Date	Speakers	TV station
11.09.2017	Strache (Freedom Party) vs. Strolz (NEOS - The New Austria)	Puls 4
11.09.2017	Strache (Freedom Party) vs. Lunacek (Green Party)	Puls 4
18.09.2017	Kern (Social Democrats) vs. Lunacek (Green Party)	Puls 4
18.09.2017	Kern (Social Democrats) vs. Strolz (NEOS - The New Austria)	Puls 4
19.09.2017	Hofer (Freedom Party) vs. Lunacek (Green Party)	ORF 2
21.09.2017	Strolz (NEOS - The New Austria) vs. Kern (Social Democrats)	ORF 2
24.09.2017	Complete Debate (All Candidates)	Puls 4
25.09.2017	Kurz (New People's Party) vs. Lunacek (Green Party)	Puls 4
25.09.2017	Kurz (New People's Party) vs. Strolz (NEOS - The New Austria)	Puls 4
26.09.2017	Kern (Social Democrats) vs. Lunacek (Green Party)	ORF 2
27.09.2017	Strache (Freedom Party) vs. Griss (NEOS - The New Austria)	ORF 2
28.09.2017	Kurz (New People's Party) vs. Lunacek (Green Party)	ORF 2
01.10.2017	Complete Debate (All Candidates)	ATV
02.10.2017	Kern (Social Democrats) vs. Strache (Freedom Party)	Puls 4
02.10.2017	Lunacek (Green Party) vs. Strolz (NEOS - The New Austria)	Puls 4
03.10.2017	Moser (New People's Party) vs. Strolz (NEOS - The New Austria)	ORF 2
05.10.2017	Strolz (NEOS - The New Austria) vs. Lunacek (Green Party)	ORF 2
08.10.2017	Kurz (New People's Party) vs. Kern (Social Democrats)	Puls 4
08.10.2017	Kurz (New People's Party) vs. Strache (Freedom Party)	Puls 4
09.10.2017	Strache (Freedom Party) vs. Kern (Social Democrats)	ORF 2
10.10.2017	Kurz (New People's Party) vs. Strache (Freedom Party)	ORF 2
11.10.2017	Kurz (New People's Party) vs. Kern (Social Democrats)	ORF 2
12.10.2017	Complete Debate (All Candidates)	ORF 2

10 Robustness Checks for Austrian Election Analysis

While we report confidence intervals based on bootstrapped standard errors in Table 2 in the paper, we here simply report the same results based on analytical OLS standard errors in Table A3. Our substantive conclusions do not change. For our analyses in Table 2 in the paper, we constructed the dataset by candidates and excluded three observations from debates when parties sent substitutes for their lead candidate. Table A4 shows that our results are robust to including these three observations. For the analyses in the paper we operationalize parties' positions as the mean placement of parties by CHES experts. In Table A5, we alternatively operationalize the position as the median placement by experts to diminish the weight of the second mode in bi-modal distributions of expert placements. The substantive results do not change.

Table A3: Explaining Campaign Positions of Party Leaders (Austria 2017)

	Model A1	Model A2	Model A3
Left-Right Position of Debating Partners	**0.05** [0.01; 0.08]		
Migration Policy Position of Debating Partners		**0.05** [0.03; 0.07]	
GAL-TAN Position of Debating Partners			**0.06** [0.04; 0.08]
Fixed effects	Parties	Parties	Parties
R^2	0.13	0.31	0.43
N	52	52	52

Note: 95% confidence intervals in parentheses.

Table A4: Explaining Campaign Positions of Party Leaders (Austria 2017)

	Model A1	Model A2	Model A3
Left-Right Position of Debating Partners	**0.09** [0.01; 0.18]		
Migration Policy Position of Debating Partners		**0.10** [0.05; 0.15]	
GAL-TAN Position of Debating Partners			**0.11** [0.07; 0.14]
Fixed effects	Parties	Parties	Parties
R^2 (within)	0.12	0.24	0.26
N	55	55	55

Note: 95% confidence intervals from nonparametric bootstrap resampling on the level of candidates (2,000 samples).

Table A5: Explaining Campaign Positions of Party Leaders (Austria 2017)

	Model A1	Model A2	Model A3
Left-Right Position of Debating Partners	**0.04** [0.03; 0.06]		
Migration Policy Position of Debating Partners		**0.05** [0.03; 0.06]	
GAL-TAN Position of Debating Partners			**0.06** [0.04; 0.07]
Fixed effects	Parties	Parties	Parties
R^2 (within)	0.13	0.29	0.41
N	52	52	52

Note: 95% confidence intervals from nonparametric bootstrap resampling on the level of candidates (2,000 samples).

11 Country Codes

Table A6: Country Codes

Code	Country
AT	Austria
BE	Belgium
BG	Bulgaria
CY	Cyprus
CZ	Czech Republic
DK	Denmark
DE	Germany
EE	Estonia
EL	Greece
ES	Spain
FI	Finland
FR	France
HR	Croatia
HU	Hungary
IE	Ireland
IT	Italia
LV	Latvia
LT	Lithuania
LU	Luxembourg
MT	Malta
NL	Netherlands
PL	Poland
PT	Portugal
RO	Romania
SI	Slovenia
SK	Slovakia
SE	Sweden
UK	United Kingdom

12 JAGS Code

Below we report the JAGS code we used for the standard Wordshoal model ("factor_model") and the dynamic formulation of the Wordshoal model ("dynamic_factor_model") in section "Application: Budget Negotiations in the Council of the EU" in the paper.

```
factor_model <- 'model{
#loop through actors
for(i in 1:nactors){
#loop through debates
for(j in 1:ndebate){
Y[i,j] ~ dnorm(mu[i,j], tau[i])
mu[i,j] <- alpha[j] + beta[j] * theta[actor[i]]
}}
#set normal priors on betas and alphas
for(j in 1:ndebate){
beta[j] ~ dnorm(0, 4)
alpha[j] ~ dnorm(0, 4)
}
#set priors on thetas (fix space with two actors)
theta[1] ~ dnorm(1, 1)
theta[2] ~ dnorm(-1, 1)
for(c in 3:nactors){
theta[c] ~ dnorm(0, 1)
}
#set prior on tau
for(i in 1:nactors){
```

```
tau[i] ~ dgamma(1, 1)
}}'



dynamic_factor_model <- 'model{
#loop through actors
for(i in 1:nactors){
#loop through debates
for(j in 1:ndebate){
Y[i,j] ~ dnorm(mu[i,j], tau[i])
mu[i,j] <- alpha[j] + beta[j] * theta[actor[i], period[j]]
}
change[i] <- theta[i,2] - theta[i,1]
}
#set normal priors on betas and alphas
for(j in 1:ndebate){
beta[j] ~ dnorm(0, 4)
alpha[j] ~ dnorm(0, 4)
}
#set priors on thetas (fix space with two actors)
theta[1, 1] ~ dnorm(1, 1)
theta[2, 1] ~ dnorm(-1, 1)
for(t in 2:nperiods){
theta[1, t] ~ dnorm(theta[1, t-1], tau.evol)
theta[2, t] ~ dnorm(theta[2, t-1], tau.evol)
}
```

```
for(c in 3:nactors){
theta[c, 1] ~ dnorm(0, 1)
for(t in 2:nperiods){
theta[c, t] ~ dnorm(theta[c, t-1], tau.evol)
}}
#set prior on tau
for(i in 1:nactors){
tau[i] ~ dgamma(1, 1)
}
#set priors on evolution parameters for thetas
tau.evol <- 1
}'
```

13 Convergence Diagnostics for JAGS Models

In the section "Application: Budget Negotiations in the Council of the EU" in the paper, we present results from a standard Wordshoal model as well as a dynamic version of this model over two periods. Here, we report convergence diagnostics for the underlying JAGS models (one chain with 1,000,000 iterations). Figure A12 displays a histogram of the Geweke statistics for our 64 parameters in the static Wordshoal model, and Figure A13 a histogram for our 90 parameters in the dynamic version of the Wordshoal model. This demonstrates that almost all Geweke statistics lie between -2 and +2. In fact, only one parameter in the static version and three parameters in the dynamic version are more extreme than these values. This demonstrates the generally good convergence of the sampler to its stationary distribution.

Figure A12: Geweke Statistics from Static Wordshoal Model

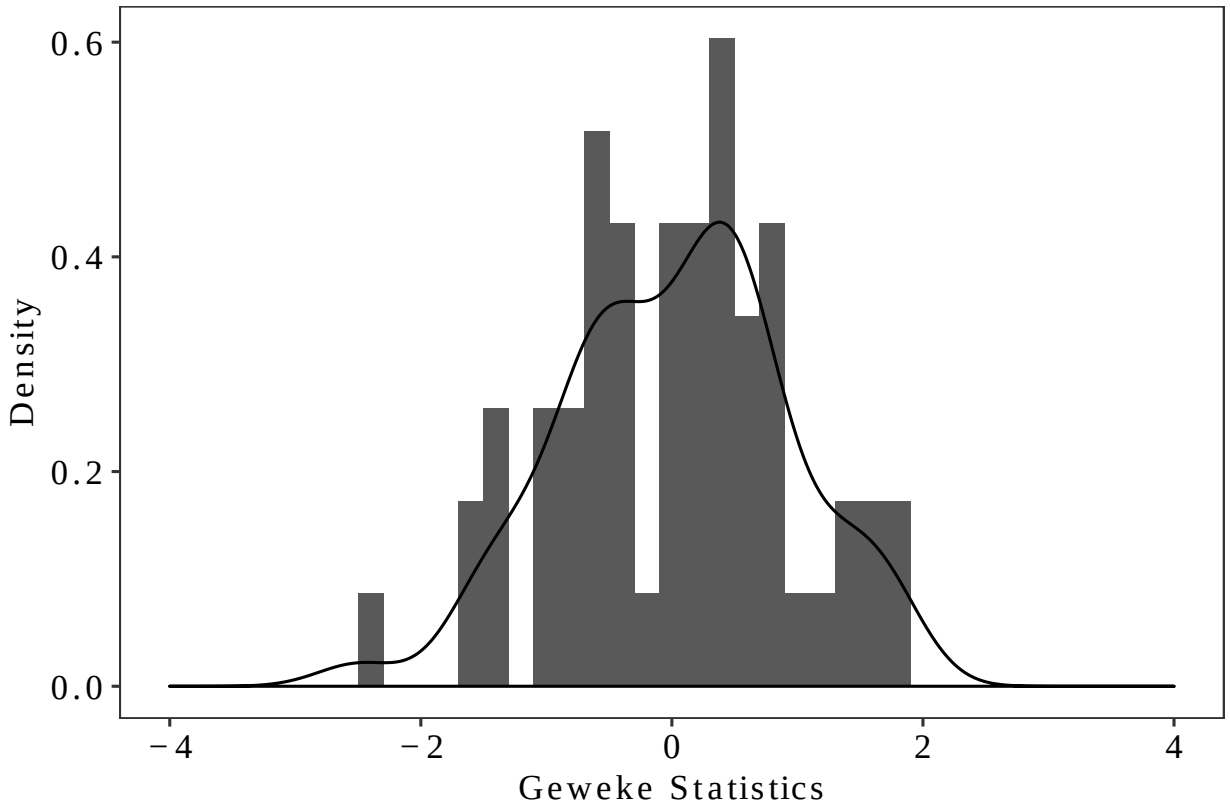
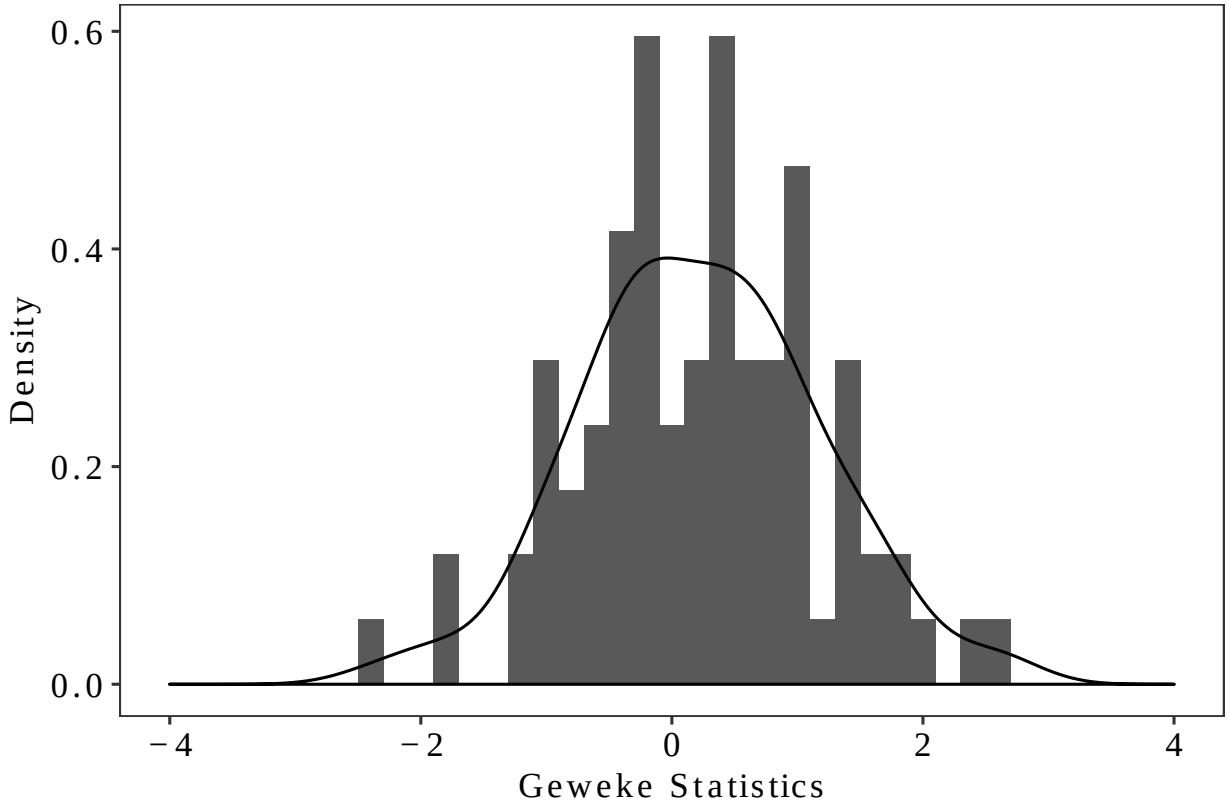


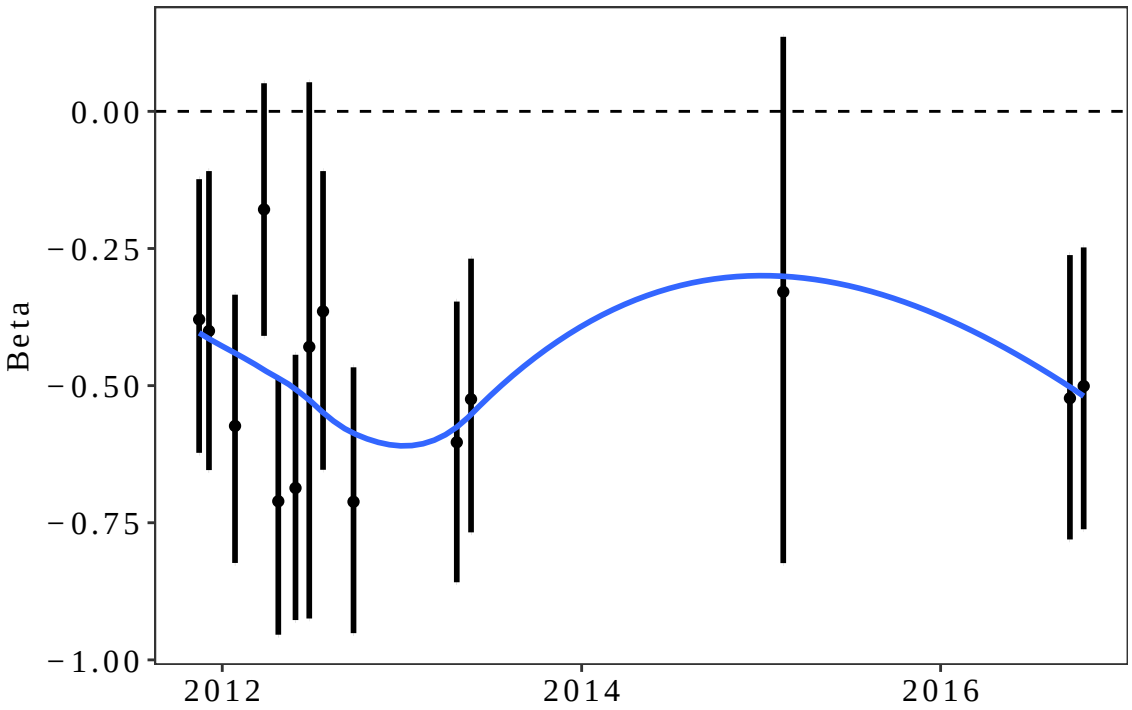
Figure A13: Geweke Statistics from Dynamic Wordshoal Model



14 Debate Loadings from Wordshoal Models

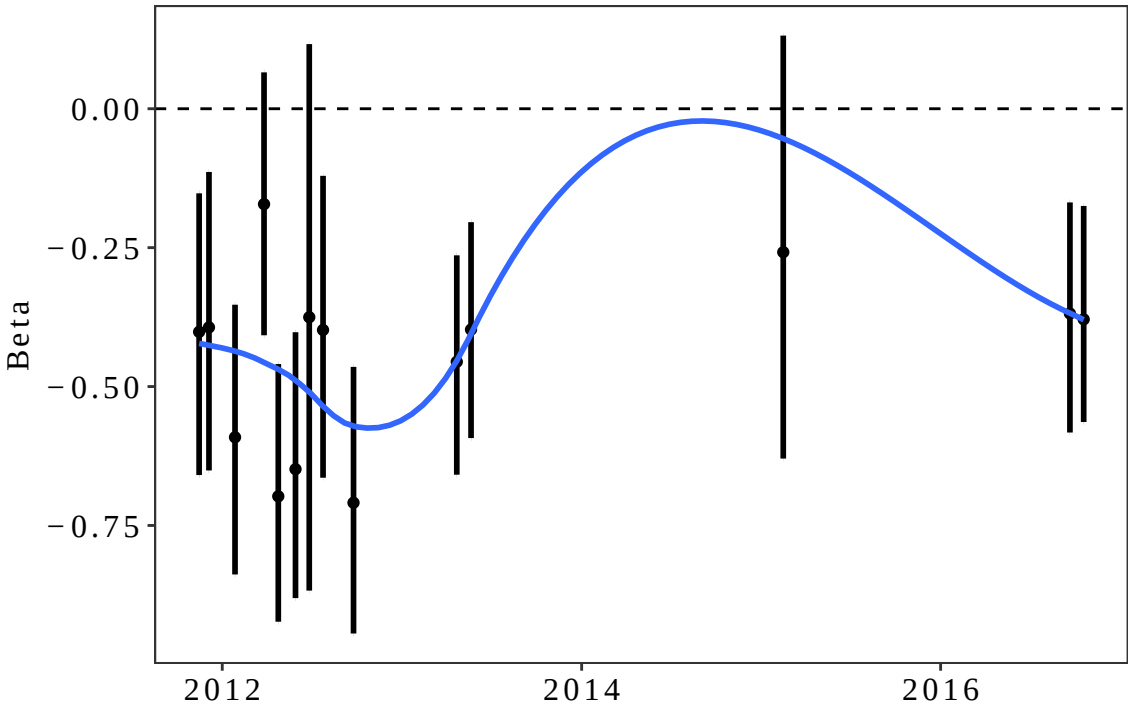
In Figures A14 and A15 we plot the debate loadings (β_j) from our static and dynamic Wordhsoal models over time. This also reveals that debate loadings (except for the first debate with very few speakers) slightly increase over time in both models, which is consistent with our finding of a polarization in the estimated positions ($\theta_{i,t}$) between period 1 and period 2 in the dynamic version of the model.

Figure A14: Debate Loadings from Static Wordshoal Model



Note: 95% credible intervals as vertical lines.

Figure A15: Debate Loadings from Dynamic Wordshoal Model

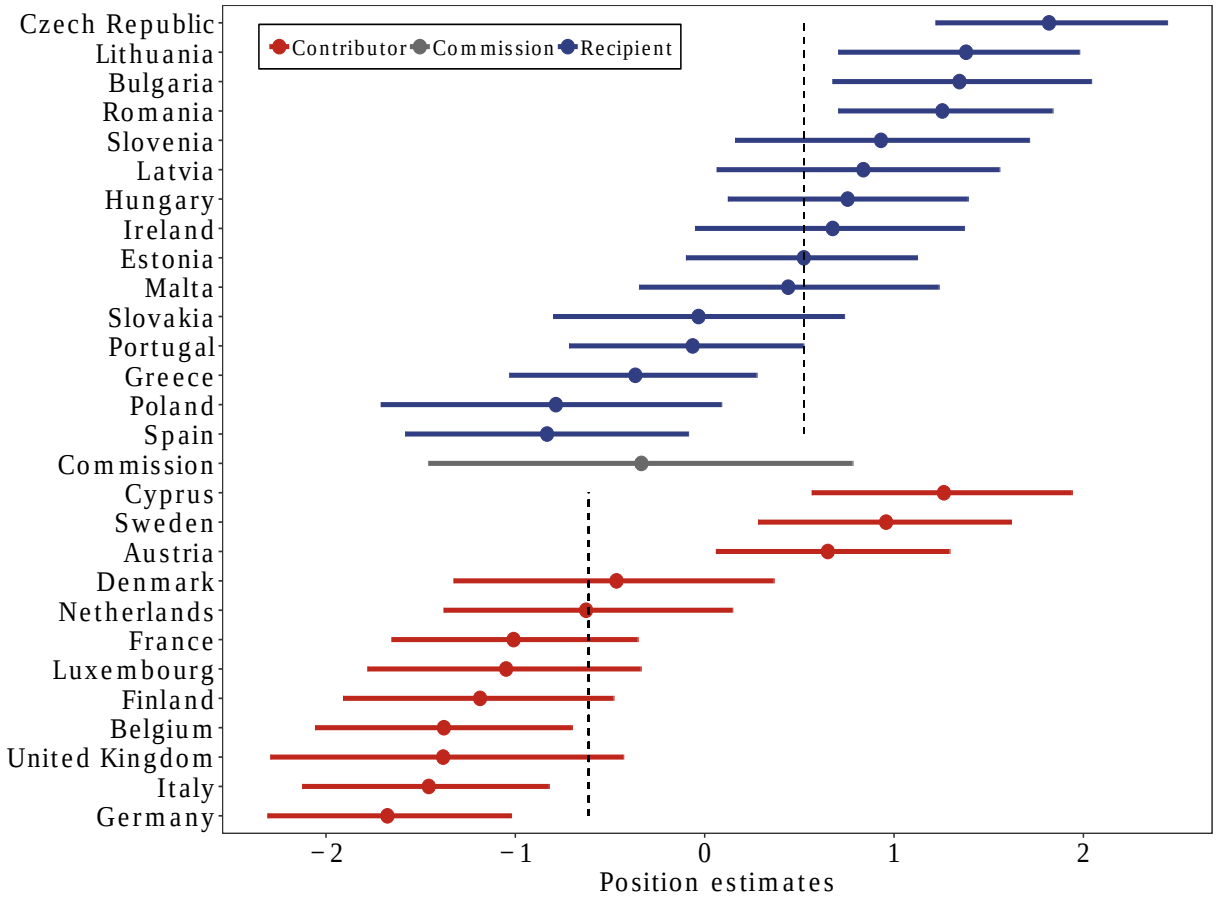


Note: 95% credible intervals as vertical lines.

15 Replicating the substantive analysis of the MFF debates with the API corpus

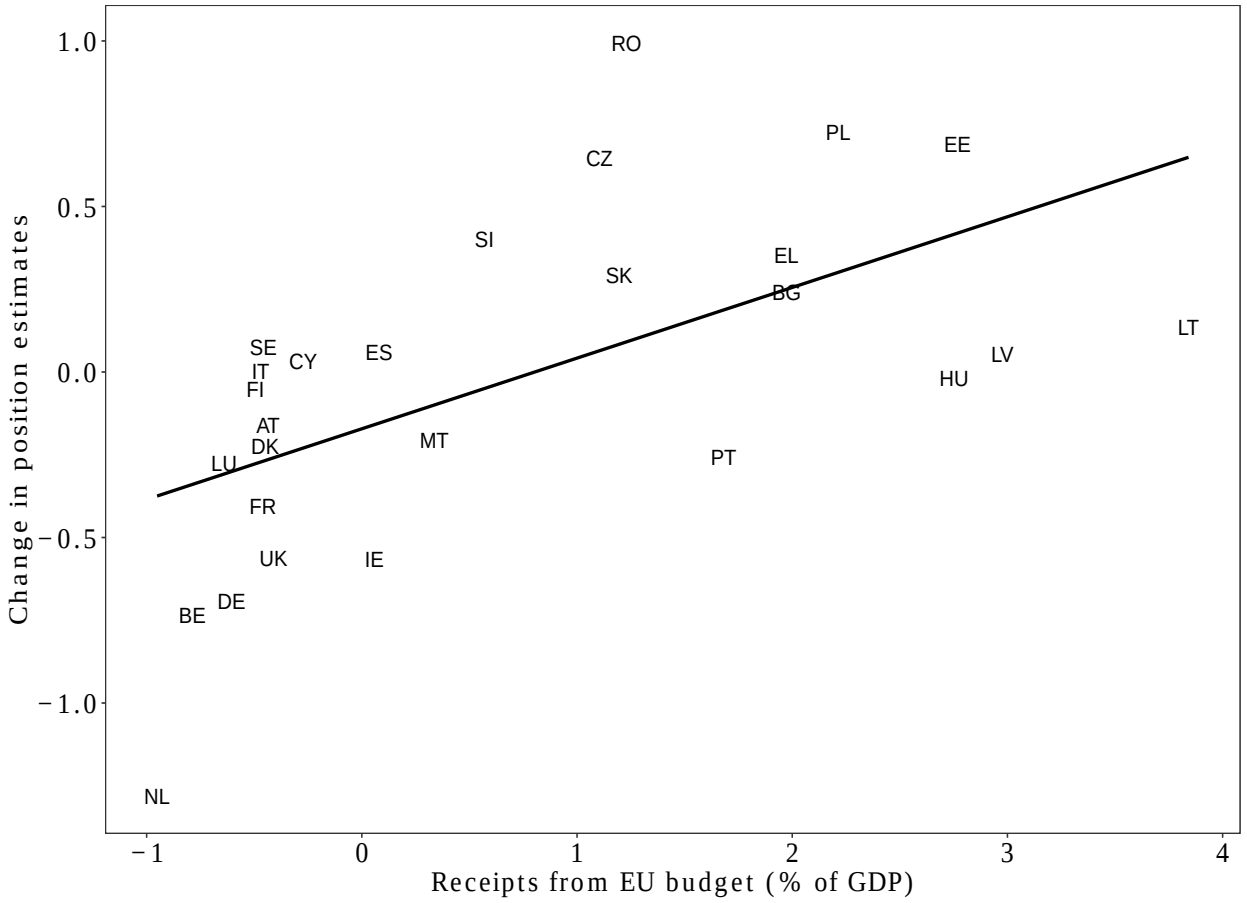
Figures A16 and A17 replicate the substantive findings in chapter 6. They show similar results. In Figure A16, the difference in means between the two groups is 1.13 (1.09 in the YouTube corpus). The correlation shown in Figure A17 is 0.59 (0.60 in the YouTube corpus).

Figure A16: Government Position Estimates in EU MFF Negotiations 2011-2016 (Wordshoal) (with API corpus)



Note: 95% credible intervals as vertical lines.

Figure A17: Relationship between Change in Position and Receipts from the EU Budget (with API corpus)

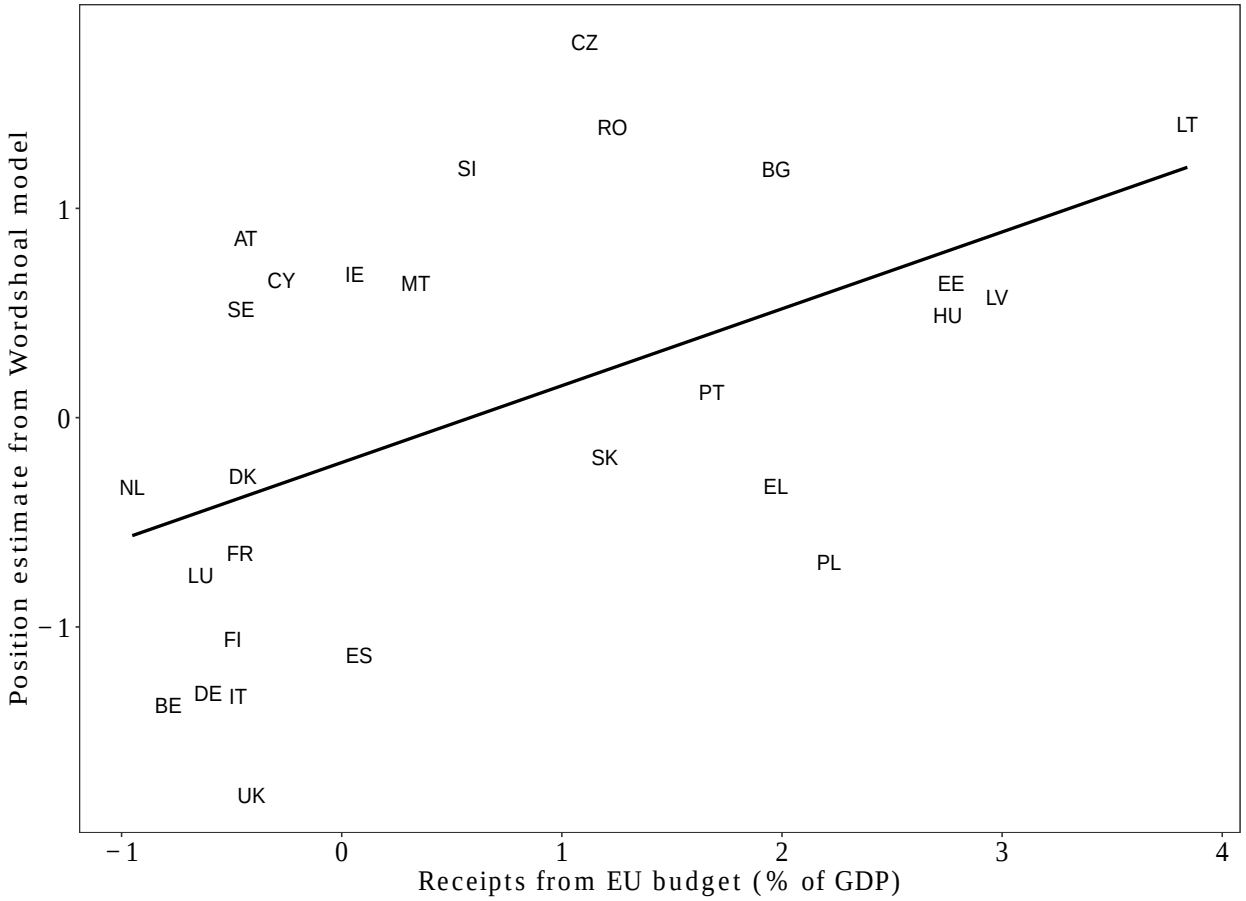


Note: 95% credible intervals as vertical lines.

16 Wordshoal Position in Relation to Receipts from EU Budget in MFF Debates

Figure A18 shows the position estimate from the Wordshoal model in relation to receipts from the EU budget (as % of GDP). The position in the debate is already correlated with the receipt from the budget, a trend that becomes even stronger during the negotiation.

Figure A18: Wordshoal Position in Relation to Contribution for MFF debates

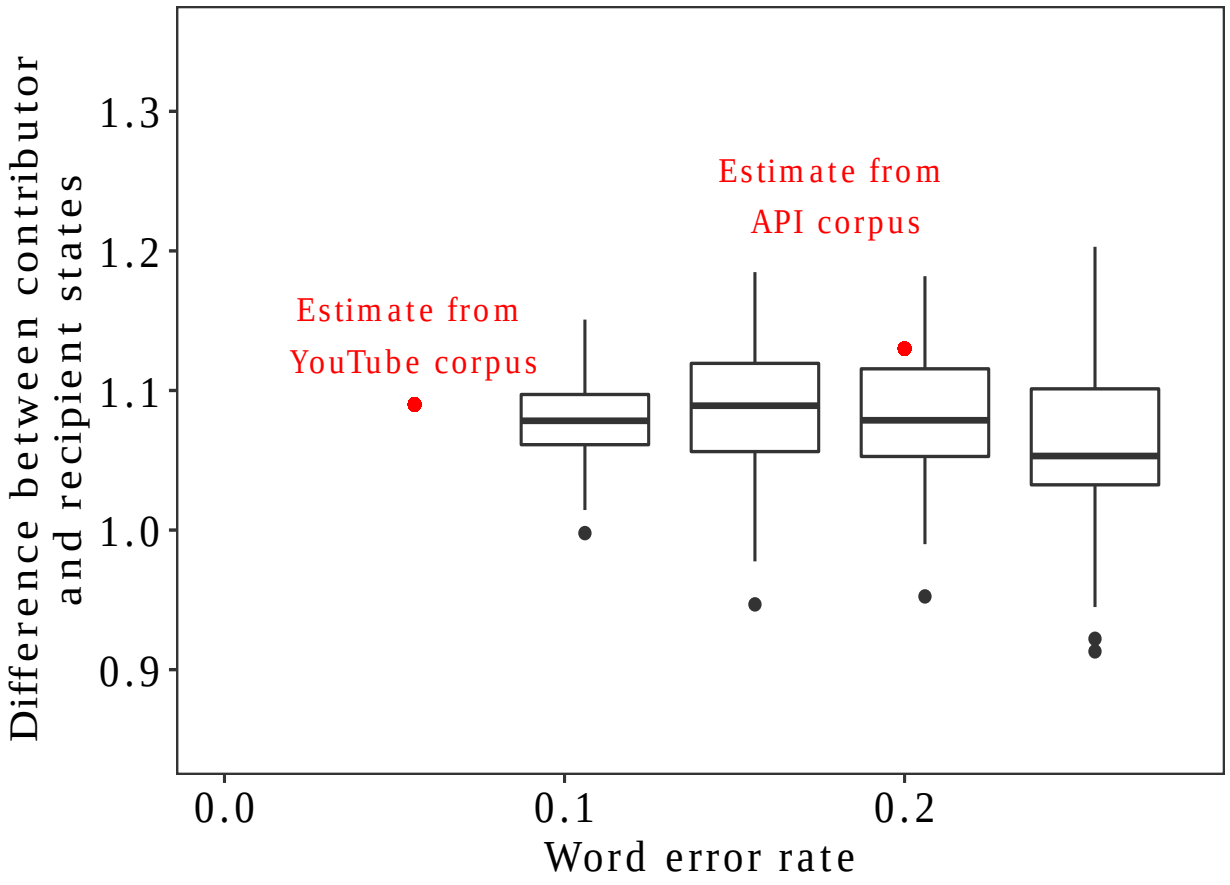


Note: 95% credible intervals as vertical lines.

17 WERSIM-simulated Difference between Contributor and Recipient States in MFF Debates

Figure A19 shows the simulated differences between contributor and recipient states in the MFF debates. The simulated quantity is the difference in means as depicted in Figure 5 in the paper. We simulated 50 corpora each for additional word error rates of 0.05, 0.1, 0.15 and 0.2 and calculated the difference of means between the groups. Even after introducing substantial amounts of additional word error, the quantity of interest is consistently positive around a value of 1.1. We also plotted the estimates from YouTube and the API. The estimate from the API have a measured Word Error Rate of 0.2 and lie within the range of simulated difference for corpora with similar word error rates.

Figure A19: Simulated Differences between Contributor and Recipient States Using WER-SIM



18 Information on Human Transcriptions

All our human transcriptions were created by trained research assistants. The assistant who produced transcriptions in English had lived in an English-speaking country before. The assistant who produced transcriptions in German was a native speaker. The assistant who produced transcriptions in French had attended a bilingual French high school. A sample of the transcriptions in English was checked by the authors for high accuracy. A sample of the transcriptions in French was checked by a French translator. In transcribing the materials the assistants adhered to the following guidelines:

- Only transcribe **whole words** with semantic content (i.e. no phonemes, no interjections).
- Transcribe any **grammar** in its correct form (e.g. apply the apostrophes correctly).
- Apply **hyphenation** as suggested by language dictionaries (e.g. "policy-making").
- Fully transcribe all **repetitions** (e.g. "I I will be making a a decision soon").
- **Numbers** are transcribed as follows:
  - Zero to twelve are spelled out, larger numbers are transcribed as numerals.
  - Numbers that make short words are also spelled out, especially round numbers: twenty, hundred, three thousand.
  - Decimals and equations are always written in numerals. Thus: "4 + 5 = 9" and "3.5". This also includes percentages.
  - Roughly estimated figures are spelled out, accurate figures are written in numerals, e.g. "The fifty million Euros in state subsidues".
  - Follow established conventions regarding spelling. Street addresses, page numbers, telephone numbers, bank account numbers, dates, headings etc. are never written out. For instance: "on page 11" or "16 Broad Street".