EUSpeech: a New Dataset of EU Elite Speeches

Gijs Schumacher
University of Amsterdam
g.schumacher@uva.nl

Martijn Schoonvelde
Vrije Universiteit, Amsterdam
h.j.m.schoonvelde@vu.nl

Denise Traber
University of Zurich
traber@ipz.uzh.ch

Tanushree Dahiya
University of Amsterdam
tanushree.dahiya@student.uva.nl

Erik de Vries
University of Amsterdam
erik@devries.pm

Abstract

This paper presents EUSpeech, a new dataset of 18,403 speeches from EU leaders (i.e., heads of government in 10 member states, EU commissioners, party leaders in the European Parliament, and ECB and IMF leaders) from 2007 to 2015. These speeches vary in sentiment, topics and ideology, allowing for fine-grained, over-time comparison of representation in the EU.

1 Introduction

This paper presents EUSpeech, a new dataset of 18,403 speeches from EU leaders (i.e., heads of government in 10 member states, EU commission members, party leaders in the European Parliament, and ECB and IMF leaders) from 2007 to 2015 (Schumacher et al., 2016).[1] These speeches vary in sentiment, topics and ideology, allowing for fine-grained, over-time comparison of representation in the EU. This paper illustrates the possibilities of working with EUSpeech for scholars interested in elite-mass interactions in the EU. To this end, the next section first introduces EUSpeech. We then present a Wordfish scaling analysis, identifying a clear anti-Europe vs pro-Europe dimension in EP speeches (Slapin and Proksch, 2008; Proksch and Slapin, 2009). Furthermore, we use sentiment analysis to show that speech sentiment responds to objective economic and political factors (Young and Soroka, 2012).

2 EUSpeech

EUSpeech consists of all publicly available speeches from the main European institutions plus the IMF and the speeches of prime ministers—or president in the case of France—of 10 EU countries for the period after 1 January 2007.[2] Most countries and institutions have a dedicated website that stores information on the decisions, background, media appearances and speeches of members of government. In most cases websites clearly demarcated speeches from other types of oral communication such as interviews or debates.[3]

Table 1 gives an overview of the institutions and countries in our dataset and the websites we collected speeches from.[4] In most cases we used the official government websites.[5] Interestingly, most official government websites delete the speeches of outgoing premiers or presidents, leaving us with

[1] This dataset is available on Harvard Dataverse: https://dataverse.harvard.edu/dataverse/euspeech.

[2] These countries are Czech Republic, France, Germany, Greece, Netherlands, Italy, Spain, United Kingdom, Poland and Portugal.

[3] We did not collect these other types of oral communication because they depend on third parties.

[4] EUSpeech also includes the Python scripts we used to scrape the speech texts and metadata.

[5] For France we found a non-governmental website that had collected all the speeches from the relevant Presidents.

	Total	English	Speakers	Source	Wayback machine	Time period
Institution						
IMF	509	509	-	imf.org	No	01/2007 - 11/2015
European Council	236	220	2	consilium.europe.eu	No	11/2009 - 09/2015
European Commission	6140	5991	-	europa.eu	No	01/2007 - 11/2015
European Central Bank	1008	990	-	ecb.europa.eu	No	01/2007 - 11/2015
European Parliament	3665	2698	26	europarl.europa.eu	No	01/2007 - 11/2015
ALDE	48	43	1	alde.eu	No	10/2010 - 11/2014
ECR	56	55	1	ecrgroup.eu	No	07/2009 - 10/2015
Country						
Czech Republic	273	39	4	vlada.cz	Yes	06/2009 - 11/2015
France	1451	0	3	vie-publique.fr	No	01/2007 - 10/2015
Germany	580	1	1	bundeskanzlerin.de	No	10/2008 - 11/2015
Greece	484	94	4	primeminister.gov.gr	Yes	10/2009 - 11/2015
Netherlands	392	132	2	rijksoverheid.nl	No	02/2007 - 11/2015
Italy	867	63	5	governo.it	Yes	01/2008 - 9/2015
Poland	4	0	3	premier.gov.pl	No	11/2011 - 11/2015
Portugal	139	6	3	portugal.gov.pt	Yes	10/2009 - 12-2015
Spain	1764	768	2	lamoncloa.gob.es	No	01/2007 - 11/2015
United Kingdom	787	787	3	gov.uk nationalarchives.gov.uk	Yes	03/2007 - 11/2015

Table 1: **Number of speeches per country, language and institution**

only the speeches of the incumbent premier or president. To solve this problem we used the Wayback Machine, allowing us to travel back to the governments' website prior to the change of government.[6] This way we were able to retrieve most speeches, although some missing speeches were unavoidable.[7] We did not collect speeches by interim prime ministers.

Table 1 also gives an overview of the number of speeches, the number of speakers and the period for which the speeches were collected for each country and institution. There is variation between countries on all of these criteria. Clearly, some countries had more changes in leadership than others.[8] Some countries have more speeches than others for at least two reasons: larger countries tend to have more speeches than smaller ones, and some countries are simply more diligent than others in keeping track of these speeches.[9]

All speeches were scraped using Python.[10] For each country, institution, and language, we saved the text of all speeches, as well as metadata like date, speaker, title and speech length in a single csv file. We also cleaned the scraped speeches, discarding sentence structure and interpunction, leaving us with term-document matrixes. This allows us to extract comparable measures of position (scaling models) and sentiment (sentiment models).[11]

In the next two sections, we illustrate how the EUSpeech data can be used for fine-grained, over-time analysis of representation in the EU, using sentiment analysis and scaling models.

3 Sentiment Analysis

3.1 Method

Sentiment analysis uses a dictionary that indicates whether words have positive or negative sentiment. We combined two dictionaries containing positive and negative sentiment scores of English words for in total 5875 words (Wilson et al., 2005; Mohammad and Turney, 2010). These

[6] `https://archive.org/web/`

[7] The Wayback Machine makes occasional snapshots of websites. In some cases there are a few months between the last snapshot and the change of government, thus leading to some gaps in the data.

[8] For some countries we were unable to find speeches from 2007 or 2008. These were probably never published online or are hiding in the dark corners of the internet.

[9] What is important here is whether the selection of speeches on the website is a random selection of speeches or whether specific speeches have been taken out. If a speech was important in signifying a certain position or sentiment of a leader it is unlikely to have been taken out. It is more likely that irrelevant speeches at say the opening of a rather irrelevant event run the risk of not being put online. Some important speeches, however, may not have appeared because the leader took an unpopular position that was later retracted. Unfortunately, this remains speculation.

[10] The cleaning scripts are available to users of EUSpeech as well and can be adjusted to suit their research goals.

[11] In results not presented in this paper, we also apply topic models (Grimmer, 2010), complexity analysis (Kincaid et al., 1975) and noun usage (Cichocka et al., 2016) to the speeches.

dictionaries assign identical sentiment scores to words that appear in both but combined they contain more words than they do separately. We first matched the words in the dictionaries with those in the term-document matrices. Then we calculated positive and negative sentiment scores for each speech by counting the number of positive or negative words and dividing by the total number of words. We do this for all 12,297 English-language speeches, and 6,106 speeches which were translated in English using *Google Translate*.

3.2 Results

Figure 1 reports results from the sentiment analysis. Figure 1a displays mean sentiment per quarter over all speeches. We draw two conclusions from figure 1a: (1) speeches contain almost 4 times more positive sentiment than negative sentiment; (2) positive sentiment drops dramatically after 2014. Figure 1b displays sentiment (positive and negative) by institute. Levels of sentiment differ per institute, but overall positive sentiment is present more than negative sentiment. On negative sentiment (top panel) Greece and the European Parliament score highest, and the European Council, Italy and European Commission score lowest. On positive sentiment (bottom panel) the institutions (EC, ECB, IMF, EU Council and EP) and Greece score lowest. It appears that, on average, the European institutions (plus IMF) communicate with less sentiment than the prime ministers. The prevalence of high negative sentiment and low positive sentiment for the case of Greece may reflect the disastrous economic developments there.

Figure 1c presents positive and negative sentiment for a selection of (better-known) speakers. Except for one speaker (Marcel de Graaff, co-president of Europe of Nations and Freedom), all speakers use on average more positive than negative sentiment. On average, the radical speakers in this sample (Tsipras, Farage, Bisky and De Graaff) deliver speeches with relatively more negative, and less positive sentiment than the other leaders. Speakers often seen as relatively technocratic politician types (e.g. Monti, Van Rompuy and Prodi) deliver speeches with relatively little (positive and negative) sentiment.

Finally, Figure 1d presents results from four different regression analyses of positive or negative sentiment in prime minister speeches and institutions speeches on quarterly GDP growth of the Eurozone and quarterly GDP growth of the respective country, and a political crisis variable as measured by the number EU council meetings in each time period.[12]. Figure 1d shows that negative sentiment in speeches from European institutes (plus IMF) shrinks with economic growth and the number of EU Council meetings. In other words, the better the economy or the more political crisis, the less negative sentiment in speeches. Our analysis of negative sentiment in prime minister speeches is similar in the sense that country GDP growth reduces negative sentiment. However, eurozone growth increases negative sentiment. This means that negative sentiment is especially high if the eurozone is growing, but the economy of the prime minister's country is shrinking. If both country and the eurozone economies are growing, these two effects should cancel each other out. Eurozone growth and the number of EU council meetings stimulate positive sentiment in the speeches by the European institutes (plus IMF). Hence, our results in the analysis of positive sentiment are the exact reverse of the results of negative sentiment. This is not the case for the PM speeches. Here we find no effect of growth. This suggests a loss aversion mechanism: more negative sentiment in response to economic decline, but no changes to positive sentiment.

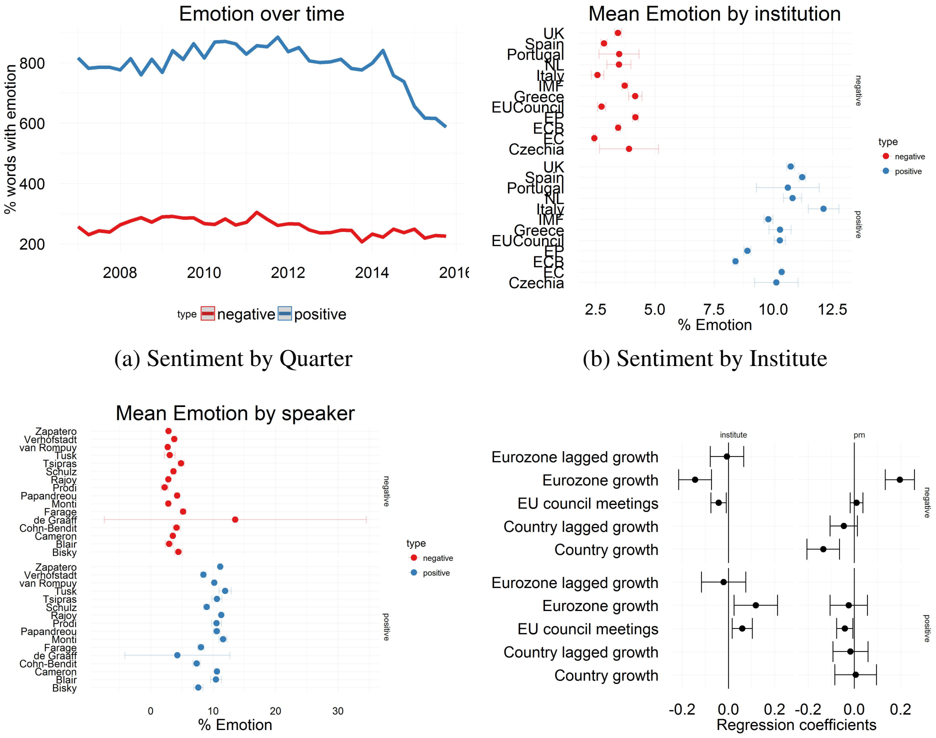
4 Scaling Models

4.1 Method

As a second illustration of the EUSpeech data we scale the European Parliament speeches for each EP group leader using *Wordfish* (Slapin and Proksch, 2008; Proksch and Slapin, 2009). *Wordfish* extracts substantively relevant quantities in an unsupervised manner, scaling these speeches on a latent dimension (Slapin and Proksch, 2008). Using select anchors (words and documents) we can retrieve the meaning of the latent dimension that *Wordfish* produces.[13] This approach relies on the assumption that the content of the political texts is predominantly ideological, and therefore informative of the policy position expressed by each actor (Grimmer and Stewart, 2013). For this analy-

[12] We also control for whether the text is translated or originally English (not presented)

[13] Exactly because we do not know *a priori* what the dominant ideological dimension is in the European Union we decided on using *Wordfish* rather than *Wordscores* which assumes we know the latent dimension.



(a) Sentiment by Quarter

(b) Sentiment by Institute

(c) Sentiment by Speaker

(d) Regression analysis Sentiment

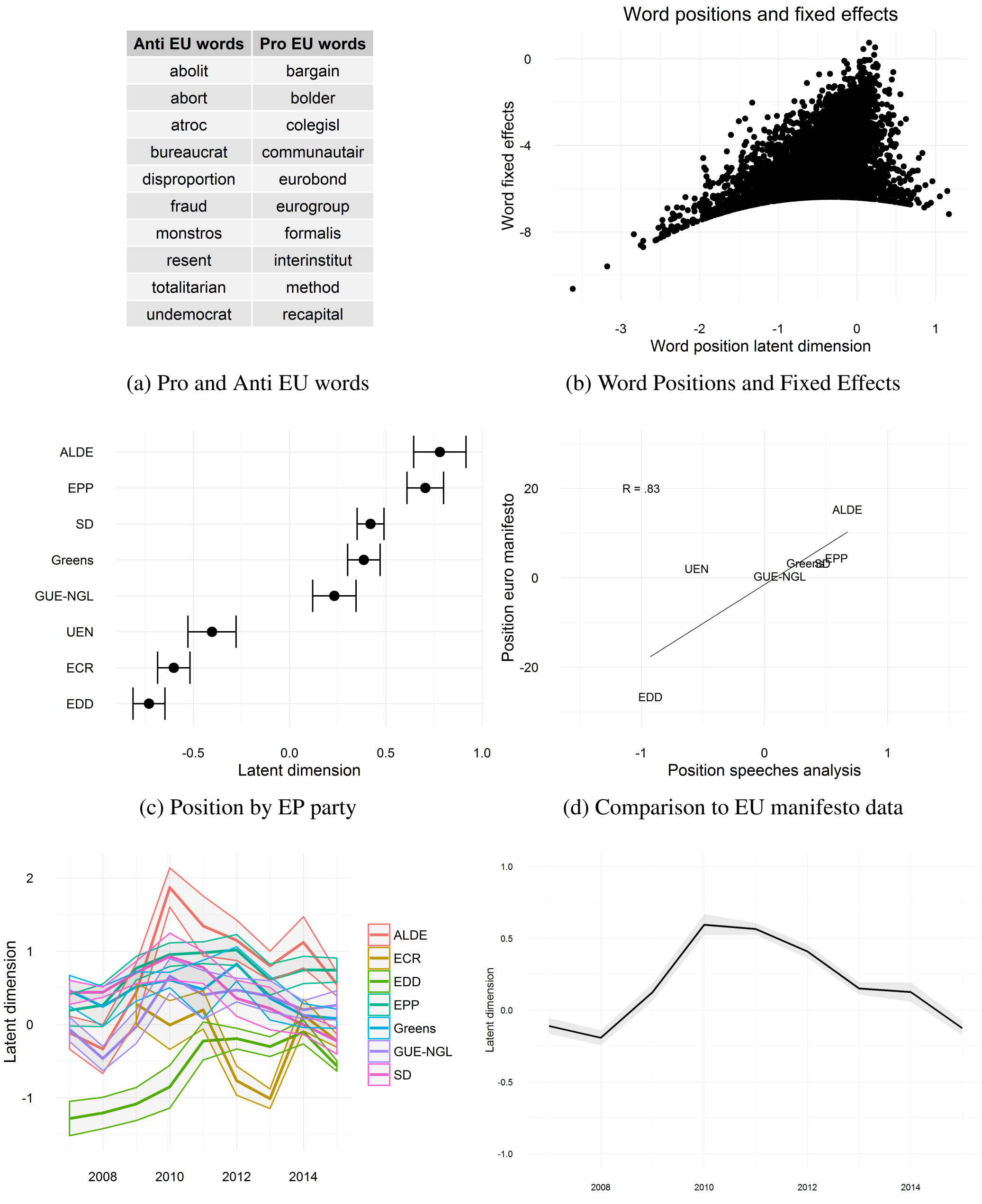
Figure 1: **Sentiment Analysis**

sis we translated the roughly 1000 non-English EP speeches using *Google Translate*.

Results

The upper left panel in figure 2 demonstrates the placement of words along the single, latent continuum that wordfish estimates on the x-axis, and the words' fixed effects on the y-axis. This figure is usually referred to as an Eiffel Tower Plot. In the middle of the x-axis there are words that occur a lot, but do not distinguish positions. On the extremes of the x-axis we find words that occur less often, but are strong indicators for distance between documents. To make sense of both dimensions figure 2a lists some of the high-scoring (high betas) words on both ends of the dimension. Figure 2b shows word placement (a dot) on the ideological dimension (x-axis) and the word fixed effect (y-axis). The latter indicates how often the word occurs. Words high on the y-axis occur often and therefore do not distinguish well between documents. A word like "house" is such a word. Other words do distinguish well between documents, because some politicians use them and others do not. Some words do not occur that often (score low on y-axis) but are only used in some documents and not in others (extreme score on x-axis). We look to these words to identify the dimension that is estimated by the wordfish procedure. On the left-hand side of figure 2b we find negative word stems such as "abolit", "abort", "undemocrat" and "totalitarian". On the right-hand side we find stems such as "colegisl", "communitarian" and "Eurobond". On the basis of this we propose to identify the latent dimension as an anti-Europe vs pro-Europe dimension. Admittedly, we present here the words that make the most sense to make this case. On both ends of the dimension we also find words that are not easily placeable on any dimension. It is likely that splitting up (parts of) speeches according to topic will increase the clarity of the estimated dimension.

The wordfish analysis also estimates positions of the speeches on the latent dimension. For each party we calculate the mean of these positions and a 95% confidence interval (see figure 2c). The anti-European parties EDD, ECR and UEN cluster on the left of our dimension. The pro-European, mainstream parties EPP, ALDE and SD cluster on the right. To further validate our findings we compare our wordfish party estimates to the Euromanifesto 2009 estimates of party positions on the



Anti EU words	Pro EU words
abolit	bargain
abort	bolder
atroc	colegisl
bureaucrat	communautair
disproportion	eurobond
fraud	eurogroup
monstros	formalis
resent	interinstitut
totalitarian	method
undemocrat	recapital

(a) Pro and Anti EU words

(b) Word Positions and Fixed Effects

(c) Position by EP party

(d) Comparison to EU manifesto data

(e) Party position over time

(f) Position over time

Figure 2: **Ideological scaling**

anti-European vs pro-European scale. Figure 2d presents this data. The correlation between the two is .83. Hence, it is quite clear that our model measures a pro versus anti-European ideological dimension.

The last two plots display time trends. Figure 2e shows the mean position of parties over time with 90% confidence intervals. As is clear, there is quite some overlap between parties. The EDD is consistently the most anti-European party. The ECR fluctuates a bit more. The ENL is also anti-European, but has been omitted from this plot, since was only recently founded. ALDE and EDD are the most pro-European parties, but especially ALDE was more in the middle of the ideological scale until 2009. We ran a regression model to explain these party position changes. One, very strong predictor of party position change is party leadership change. The shift by ALDE coincides with the transition from Graham Watson to Guy Verhofstadt as party leader. Party leader changes also explain the ECR shifts. Interestingly, appointing a leader from the UK brings about a shift towards a more anti-European position. The somewhat dramatic changes to occur due to a leadership change, also suggests that leaders in European parties do not really take the middle ground of their party MEPs position. Otherwise, the party position would be more stable over time.

The final question is: what is the time trend? For this purpose we took the (unweighted) average per year of the party positions. Figure 2f displays this time trend. Initially, we see a shift towards a more pro-European position. This is primarily caused by the appointment of Verhofstadt as the ALDE leader, and by the moderation of the EDD. But after 2011 there is towards the middle of the ideological scale, towards a more euro-skeptical position. Here it is primarily ALDE and SD that moderated their pro-European position and the emergence of the ENL that shifts the mean. But also the Greens, EDD and ECR shift to a more anti-EU position.

5 Conclusion

In this paper we presented EUSpeech, a new dataset of 18,403 speeches of EU leaders, containing variation in sentiment and ideology, allowing for fine-grained analysis of representation the European Union. In analyses not presented here we also find interesting and predictable variation in speech topics, speech complexity and speech word usage. With these findings in mind, we think that EUSpeech will be a valuable resource for scholars interested in elite-mass interactions in the European Union.

6 Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 649281, EUENGAGE.

References

Aleksandra Cichocka, Michal Bilewicz, John T Jost, Natasza Marroush, and Marta Witkowska. 2016. On the grammar of politics—or why conservatives prefer nouns. *Political Psychology*, xx(xx):xx–xx.

Justin Grimmer 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.

Justin Grimmer. 2010. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35.

J.P. Kincaid, R.P. Fishburne, R.L. Rogers, and B.S. Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Memphis, Tennessee: Naval Air Station.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. *CAAGET '10 Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34.

S.-O. Proksch and J B Slapin. 2009. How to Avoid Pitfalls in Statistical Analysis of Political Texts: The Case of Germany. *German Politics*, 18(3):323–344.

Gijs Schumacher, Martijn Schoonvelde, Tanushree Dahiya, and Erik De Vries. 2016. Euspeech. dx.doi.org/10.7910/DVN/XPCVEI, Harvard Dataverse, V1.

J B Slapin 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.

T Wilson, J Wiebe, and P Hoffman. 2005. Recognizing contextual polarity in phrase level sentiment analysis. *Acl*, 7(5):12–21.

Lori Young and Stuart Soroka. 2012. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.