EUSpeech v2.0*
A dataset of 11,466 speeches by European Leaders

Gijs Schumacher[1], Nicolai Berk[1], Christian Pipal[1], Jaroslaw Kantorowicz[4], Martijn Schoonvelde[5], Denise Traber[6], and Erik de Vries[7]

[1]Department of Political Science, University of Amsterdam
[4]Institute of Security and Global Affairs, University of Leiden
[5]Department of Political Science, University College Dublin
[6]Department of Social Sciences, University of Basel
[7]Department of Media and Social Sciences, University of Stavanger

1 Introduction

Here we describe EUSpeech version 2.0. EUSpeech v2.0 is a dataset of speeches from heads of government from 14 different European countries. For each speech in the dataset we saved the original text, the length, the date, the speaker, the country of the speaker, the language and the title. We only removed html code from the original, scraped text. Compared to EUSpeech v1.0 (Schumacher et al. 2016a; Schumacher et al. 2016b) we extended the number of countries from 10 to 14 and have updated the timelines to April 2020 the latest. Also, in EUSpeech v2.0 we have only retained the

*Project design and management: G.S, data collection: N.B, J.K, data validation: C.P, N.B, G.S, write paper: G.S., N.B, comment on paper: C.P, data collection and management EUSpeech v1.0: M.S, D.T, E.V. We acknowledge funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 649281, EUENGAGE and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 759079, POLEMIC.

speeches by heads of government. Speeches from the European Parliament, European Commission, International Monetary Fund and European Central Bank are no longer collected, but are each easily retrievable from the websites of these institutions. In most countries we collected prime minister speeches because they are the head of government, and thereby they represent the most important political office in the country. The main exceptions in our data collection are Poland and France which both have a prime minister and a powerful Presidency. For that reason we collected speeches from the Polish and French presidents, and not the prime ministers.

2 Description of data collection

Our procedures to scrape the speeches can be found here: `https://github.com/nicolaiberk/EUSpeech`. There we also provide a detailed explanation of how to proceed with scraping speeches. In the next sections we only present a few of the most important details.

All speeches have been scraped from government websites or historical archives of government websites. The data collection thus consists of speeches published by the government. Governments typically do not keep archives of previous prime minister speeches. Therefore, it is difficult to obtain historical speeches. Also, in some cases we encountered difficulties in scraping government websites.

The file "SourceLinks_2.csv" contains all the links we have used to scrape the speeches. To collect these links we used the `Python` programming language[1]. The html code for most of these pages was collected using the `request` module from the `urllib` library[2]. Some pages had a more complex structure, necessitating the use of the `selenium` library to scroll and/or click buttons on the page [3]. Using xpaths to refer to the relevant html elements, we collected the title of a speech, the date it was held, and the link referring to the speech itself.

Note that some of these links are already expired or the contents have been moved. Therefore one cannot replicate the data collection.

The speeches themselves were collected from the links collected, again using the `urllib` library. Xpaths were specified to avoid collection of html code beyond the speech itself. The speeches were stripped from tabs and linebreaks, as well as trailing spaces. Sometimes collected speeches have

[1] https://www.python.org/

[2] https://docs.python.org/3/library/urllib.request.htm

[3] https://selenium-python.readthedocs.io/

been selected based on matching regular expressions[4], as some websites also contained links to interviews and/or press statements.

Apart from the cleaning steps described in the speech collection, some additional processing took place to bring the speeches into their final form. Remaining clutter from html code was removed using xpaths, and the lengths of the speech in characters and words was calculated. Languages were detected with the langdetect package for python[5].

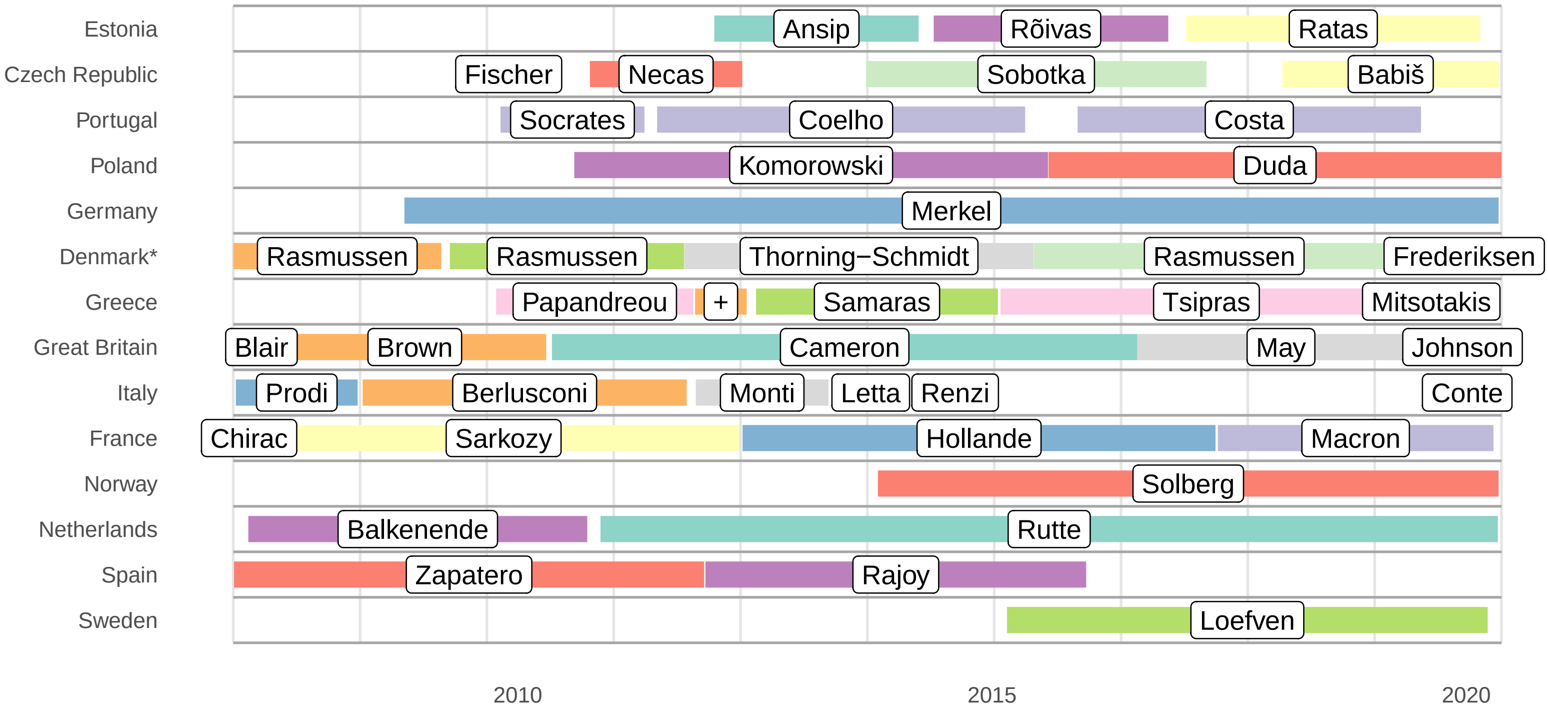
3 Description of the data

Fig 1 provides an overview of the prime ministers and presidents in our dataset. We have the longest timeline for Denmark. Our data stretches back to 1997 (cut short in the figure for plotting reasons). Denmark, Sweden, Estonia and Norway are new countries in the dataset. For Poland we switched to collecting speeches from the President. These were also widely available, in contrast to the Prime Minister speeches. We had to discontinue our data collection for Spain. Speeches are published on the governments' website but so far they have resisted our efforts to be scraped. The Italian speeches were already a challenge in the data collection of EUSpeech v1. This time we were unable to retrieve speeches from Paolo Gentiloni.

[4] https://docs.python.org/3/library/re.html

[5] https://pypi.org/project/langdetect/

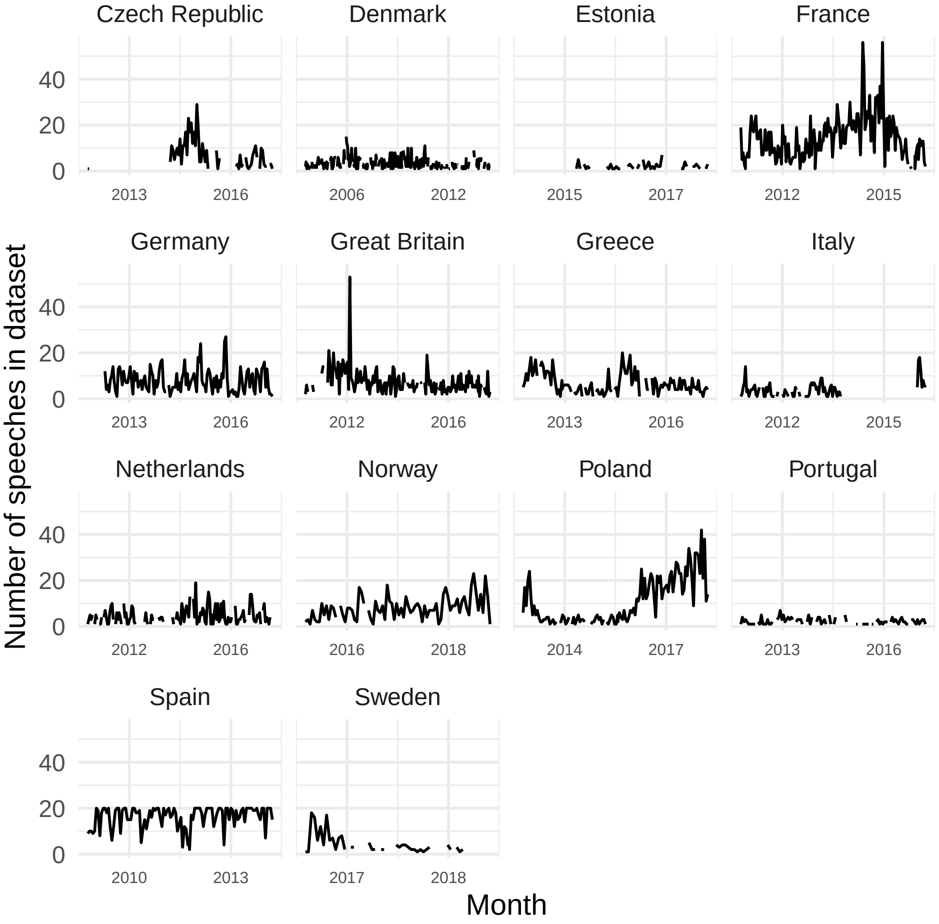
Figure 1: Leader coverage per country over time



Note: * = for Denmark the timeseries starts in 1998; + = Lucas Papademos

Fig 2 displays the frequency of speeches per month per country. Clearly there is a lot of variation between countries. The Spanish Prime Ministers and French Presidents give a lot of speeches on average. Prime Ministers from smaller countries like Estonia, Denmark, Netherlands, Portugal and Sweden give far fewer speeches. There are also several gaps in the timelines. These emerge in transition periods between governments or holidays. In the cases of Estonia and Sweden there are relatively long time gaps with no speeches. It is unclear whether this indicates that no speech was given, or no speech was published.

Figure 2: Frequency of speeches per country and over time



References

Schumacher, Gijs, Martijn Schoonvelde, Tanushree Goyal, and Erik de Vries. 2016a. *EUSpeech. A Dataset of 18,403 Speeches by EU Elites.* doi:`doi:10.7910/DVN/XPCVEI`.

Schumacher, Gijs, Martijn Schoonvelde, Denise Traber, Tanushree Dahiya, and Erik de Vries. 2016b. "EUSpeech: a New Dataset of EU Elite Speeches". In *Proceedings of the International Conference on the Advances in Computational Analysis of Political Text*, 75–80. Dubrovnik: PolText 2016. `takelab.fer.hr/poltext2016/PolText2016-proceedings.pdf`.