Replication Data for: Proksch, Lowe, Wäckerle and Soroka, 2018, Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches. Legislative Studies Quarterly

Preface

The sections below describe what the R code provided in the replication files will do. You can skip most dataset parts of the analysis and, provided you downloaded the replication files in full, the code should run. In some instances, you need to download external data. These instances are indicated below. Additionally, the replication file already comes with R datasets that let you skip the files generating these datasets. This will save you a lot of time, specifically for the ParlSpeech analysis. Essentially, you can either reproduce all steps of the analysis, or you can skip some steps like corpus construction and sentiment calculation and go right to the models and graphs.

1_SOTEU_validation

This R script runs the sentiment analysis on the State of the Union Debate in the European Parliament in 2010. It uses the automatically translated dictionaries that are stored in the RData files "auto_dictionaries_lsd.RData", lsde_frenche_germane.RData" and "auto+ictionaries_sentiws.RData". It also requires downloading the French Lexicdoder dictionary from Duval and Petry (https://www.poltext.org/en/donnees-et-analyses/lexicoder; accessed: August 2, 2018)

Duval, D. and Pétry, F. (2016) 'L'analyse automatisée du ton médiatique : construction et utilisation de la version française du Lexicoder Sentiment Dictionary', Canadian Journal of Political Science, 49(2), pp. 197–220.

The script also adds metadata to the speeches such as parliamentary group, name and country of origin, with the file "1_metadata_2010.csv"

It also reads in the human coding results, as described in the main paper (1_coding_a_2010.csv, 1_coding_b_2010.csv, 1_coding_c_2010.csv)

The script produces: Figure 1, Figure 2, Figure 3, Table A1 (Appendix) and Figure A1 (Appendix).

1_soteu_2010_protocols.RData

This is the corpus of the transcripts of all speeches during the 2010 State of the European Union debate in the European Parliament. The speeches are provided by the European Parliament on their homepage (http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+CRE+20100907+ITEM-004+DOC+XML+V0//EN; accessed: August 2, 2018) in 21 languages: Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Latvian, Lithuanian, Dutch, Polish, Portuguese, Romanian, Slovak, Slovenian and Swedish. The transcripts are then separated by speakers and are provided as separate quanteda corpora.

2_Irishbudgets.R

This R Script runs the analysis of Irish budget debates presented in chapter 4.1 "Government Dissent during Economic Crisis in Ireland" It produces: Figure 4 and Figure A2 (Appendix). As an input, it uses the Irish budget debates as collected by Will Lowe and Kenneth Benoit (2013) and combined in a quanteda corpus ("2_ireland_budgets.RData"). As dictionary for the sentiment analysis it uses the Lexicoder Dictionary, as provided in the quanteda package.

Lowe, Will and Kenneth Benoit. 2013. "Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark." Political Analysis 21(3):298-313.

3.1_Wordscores_generate_positions.R

This code and the following RData files replicate and extend the analysis of Irish budget debates that is presented in chapter 4.1. They are mainly based on the code by Herzog and Benoit (2015). For the analysis to run, we have to switch to an older version of quanteda, the code is in the R file. To get the old version to work, you might need to quit the R session and start a new session. The RScript produces a RData file called "positions_before_senti.RData" that stores the Wordscores estiamtes for later use.

Herzog, Alexander and Kenneth Benoit. 2015. "The most unkindest cuts: speaker selection and expressed government dissent during economic crisis." Journal of Politics 77(4):1157–1175.

For this and the following code to run, you need to download the following replication files for the Herzog and Benoit (2015) article store them in the replication folder. The files can be found at:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DGXXH9, version 1.0; accessed August 2, 2018.

- Data_budget_debates_1983-2012 .tab
- Data_constituency_county_link_file.tab
- Data_constituency_data.tab
- Data_county_census_population.tab
- Data_elections_data.tab
- Data_finance_ministers_1983-2013.tab
- Data_live_register_data_1983-2013.tab
- Data_members_first_Dail.tab
- Data_ministers.csv
- Data_opposition_spokesperson_1983-2013.tab

3.2_Wordscores_generate_sentiment.R

This file generates the sentiment estimates for the budget debates. To replicate our sentikment analysis for the Herzog and Benoit data, you have to use version 1.1.1 of quanteda. The code is in the R file (this is also the reason why the position and sentiment generation is split up in several files). To get this new version to work, you might need to quit the R session and start a new session. As dictionary for the sentiment analysis it uses

the Lexicoder Dictionary, as provided in the quanteda package. The RScript produces a RData file called “positions_after_senti.RData” that stores all data for later use.

3.3_Wordscores_combine.R

This file combines sentiment and Wordscores estimates. For this and the following analysis, we switch back to an older version of quanteda, the code is in the R file. To get the old version to work, you might need to quit the R session and start a new session. The RScript produces a RData file called “yearspeaker.RData” that will form the dataset for future analysis.

3.4_Wordscores_master_data.R

This script follows the analysis done by Herzog and Benoit (2015), adding data to the position and sentiment estimates. It creates a file “master_data.RData” for future use.

3.5_Wordscores_data_for_model_estimation.R
This script follows the analysis done by Herzog and Benoit (2015) and prepares the data frames for estimation. It creates R Datasets for both sentiment and Wordscores analysis.

3.6_Wordscores_estimate_models_ML.R

This script runs all models and saves them as RData files, separately for the three models (no interactions, two-way interactions, three-way interactions) and Wordscores and sentiment.

3.7_Wordscores_tables_and_coefplots.R

This code replicates: Figure 6 and Figure A4 (Appendix) based on the models run in 3.6. It also includes the code for Table A2 (Appendix) and A3 (Appendix).

3.8_Wordscores_positions.R

This code replicates: Figure 5 and Figure A3 (Appendix) and Figure A5 (Appendix) using the position and sentiment estimates saved before.

4.1_Germany_dataset.R

This script reproduces the dataset used for analysis in chapter 4.2 “Debate Sentiment and the Legislative Process in Germany”. The analysis is done on speeches on government bills given in the German Bundestag between 2005 and 2017, which were scraped from the Bundestag archive by the authors. Additionally, we added metadata on the bills these speeches were discussing. The speeches themselves are stored in the file “4_speeches_Germany.RData” and are saved as a data frame. The metadata is saved in separate files:

- 4_16leg_metadata_speakers_complete_df_encoded.Rdata
- 4_17leg_metadata_speakers_complete_df_encoded.Rdata
- 4_18leg_metadata_speakers_complete_df_encoded.Rdata
- 4_16leg_metadata _df.Rdata

- 4_17leg_metadata _df.Rdata
- 4_18leg_metadata _df.Rdata

The code uses the extended German lexicoder dictionary, which is loaded from the Rdata file "lsde_frenche_germane.RData".

The code produces a R dataset called "4_Germany_datasets.RData" for further analysis.

4.2_Germany_Plots_Models.R

This script runs the models presented in the paper and produces: Figure 7, Figure A6 (Appendix) and Table 1. It uses the 4_Germany_datasets.Rdata file produced in 4.1. It also runs Zelig models to produce estimates of predicted probabilities of unanimity as presented at the end of the chapter.

5.1_ParlSpeech_Datasets.R

This script creates the datasets that are analyzed in subsequent steps. As input, it uses the speech corpora provided on the ParlSpeech database in Version 1.0 that you need to download into the replication folder.
(https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/E4RSP9)

- Corp_Bundestag.Rdata (Germany)
- Corp_Congresso.Rdata (Spain)
- Corp_Eduskundta.RData (Finland)
- Corp_HouseOfCommons.Rdata (UK)
- Corp_PSP.Rdata (Czech Republic)
- Corp_Riksdag.Rdata (Sweden)
- Corp_TweedeKamer.Rdata (Netherlands)

Rauh, Christian; De Wilde, Pieter; Schwalbach, Jan, 2017, "The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states", https://doi.org/10.7910/DVN/E4RSP9, Harvard Dataverse, V1

The script also uses the file "5_EU_minority_govs.dta" which gives information on the minority status of governments in the countries as provided by Christopher Wratil (2016)

Wratil, Christopher. 2016. Governing Europe for the People? Citizen Representation in European Union Policy-Making. PhD Thesis, London School of Economics.

The script produces corpora for each country. As the datasets are quite large, this script takes quite a long time to run, depending on the computing power at hand.

5.2_ ParlSpeech _Sentiment.R

This script runs the sentiment analysis on the corpora created in 5.1. As input, it uses the automatically translated Lexicoder dictionaries and the extended English version of the dictionary. These are loaded from the files "auto_dictionaries_lsd.RData" and "lsde_frenche_germane.RData".

The script produces datasets for each country. As the datasets are quite large, this script takes quite a long time to run, depending on the computing power at hand.

5.3_ ParlSpeech_Models_graphs.R

This script adds in positional estimates for the parties and combines all countries. It uses the 1999-2014 Chapel Hill Expert Survey (CHES) trend file, available at (https://www.chesdata.eu/1999-2014-chapel-hill-expert-survey-ches-trend-file; accessed August 2, 2018)

Polk, Jonathan, Jan Rovny, Ryan Bakker, Erica Edwards, Liesbet Hooghe, Seth Jolly, Jelle Koedam, Filip Kostelka, Gary Marks, Gijs Schumacher, Marco Steenbergen, Milada Vachudova and Marko 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 (January-March): 1-9.
and
Ryan Bakker, Catherine de Vries, Erica Edwards, Liesbet Hooghe, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Vachudova. 2015."Measuring party positions in Europe: The Chapel Hill expert survey trend file, 1999-2010." Party Politics 21.1: 143-152.

Then it runs the model presented in the paper. Finally, it reproduces: Figure 8, Figure 9 and Table A4 (Appendix).