Corpus of Parliamentary Debates of the National Assembly of Serbia, 1997-2020

Michal Mochtak, Josip Glaurdić, and Christophe Lesschaeve

1 April 2022 (v1.0)

Contents

Overview 1

Citation Information 1

Acknowledgement 2

Variables 2

id . . . 2
term_id . . . 2
term1 . . . 2
term2 . . . 2
date . . . 2
meeting . . . 2
meeting2 . . . 2
agenda . . . 2
tag . . . 2
moderator . . . 3
fullname . . . 3
party . . . 3
party_proxy . . . 3
speech_link . . . 3
speech . . . 3
lem . . . 3
codemp . . . 4
codeparty . . . 4

Overview

This is a codebook for the corpus of parliamentary debates of the National Assembly of Serbia collected from the website Otvoreni parlament [link]. The corpus covers the period of 1997-2020 and counts eight terms and over 300 thousand speeches. As spreadsheet software (e.g. MS Excel, Libre Office Calc) is limited in how many characters can be stored in a cell, corpus data are made available in R's native binary data format "RDS".

Citation Information

If you use the dataset, please cite: Mochtak, Michal, Josip Glaurdić, and Christophe Lesschaeve (2022): SRBCorp: Corpus of Parliamentary Debates in Serbia (*vX.X*), https://doi.org/10.5281/zenodo.6521649.

Acknowledgement

The creation of the corpus was supported by the European Research Council Starting Grant [#714589]. We want to thank Leo Fel for his assistance with missing data entries in the database of MPs in Serbia.

Variables

The dataset contains 18 variables. The following overview presents all of them. Each variable is accompanied by a data type - [*character*] for a string of text; [*numeric*] for numbers; and date format for dates e.g. [*yyyy*].

id

Unique ID in the whole corpus (across terms). [*numeric*]

term_id

Unique ID per term. [*numeric*]

term1

Term duration. [*yyyy-yyyy*]

term2

Official numeric denominator for consecutive terms. The corpus covers eight terms (4th - 11th). [*numeric*]

date

Date of a speech [date format: *yyyymmdd*]

meeting

Number assigned to a meeting. As Otvoreni parliament does not store this information consistently across terms, *meeting2* may contain additional information regarding the official taxonomy. [*numeric/character*]

meeting2

Additional details on *meeting.* [*character*]

agenda

Full description of agenda (if available). Otvoreni parlament does not systematically map the agenda points. Only the most important debates are flagged. Therefore, this should be used with caution. Many of the agenda points (especially in the earlier terms) are announced by the debate's moderator and then passed without an actual discussion. [*character*]

tag

Policy category applied to an agenda point. The 21-categories coding scheme comes from the methodology of Comparative Agendas Project and related applications (e.g. Croatian Parliament). The agenda points are coded using ML model trained on known tags of agenda points from the parliaments of Croatia and Bosnia-Herzegovina. Coding scheme: (1) Bankarstvo, financije i domaća trgovina [Banking, finance, and domestic trade]; (2) Državno zemljište, upravljanje vodama i teritorijalna pitanja [State land, water management, and territorial affairs]; (3) Energija [Energy]; (4) Imigracija i izbjeglice [Immigration and refugees]; (5) Kultura i sport [Culture and sport]; (6) Ljudska prava, manjinska pitanja i građanske slobode [Human rights, minority affaris, and civic rights]; (7) Međunarodni odnosi i međunarodna pomoć [International relations and international aid]; (8) Obrana [Defense]; (9) Obrazovanje [Education]; (10) Poljoprivreda [Agriculture]; (11) Poslovi vlasti [Government affairs]; (12) Pravosuđe, kriminal i obiteljska pitanja [Justice, crime, and family affairs]; (13) Promet [Transportation]; (14) Rad i zapošljavanje [Labor and employment]; (15) Razvoj zajednice i stambena pitanja [Development and housing]; (16) Socijalne politike [Social policy]; (17) Unutarnja makroekonomska pitanja [Domestic macroeconomic affairs]; (18) Vanjska trgovina [Foreign trade]; (19) Zaštita okoliša [Environmental protection]; (20) Zdravlje [Health]; (21) Znanost, tehnologija i komunikacije [Science, technology, and communications]. [*character*]

moderator

A dummy variable for the moderator: (1) moderator; (0) others. [*numeric*]

fullname

Full name of the speaker. *last name, first name* format. [*character*]

party

Party affiliation of the speaker. Otvoreni parlament does not systematically map the party affiliation of the MPs and sometimes even provides contradictory information (e.g. two fields indicating different affiliations). We decided to keep the in-debate references to party affiliation if available for the sake of consistency and time/place binding reference, but it is recommended to use a separate MPs dataset for extracting partisanship on the individual level (see *codemp*). This is especially relevant for the earlier terms where the issue is more common. [*character*]

party_proxy

Party affiliation used for linking the entries in corpus with the database on political parties. Otvoreni parlament is unfortunately not very consistent when it comes to listing partisanship which leads to problems with linking the listed affiliations with parties in the database. This is particularly problematic when some MPs may “represent” their political party as well as the coalition they ran for or when there is a new political formation that did not run in the last election. In order to mitigate this problem, party_proxy is a party link that can be used for connecting the corpus instances with the party database (via *codeparty*). If the party exists in the party database, its party_proxy does not change. If it does not exist in the database, a proxy_party is chosen instead. This may be a result of various inconsistencies in party names, establishment of a new political subject (party splits), or listing a coalition instead of political party. In order to further improve the proxy link, the missing entries are further populated via declared party affiliation on the level of individual MPs. [*character*]

speech_link

Link to the transcript of parliamentary speeches. [*character*]

speech

Raw transcript of parliamentary speeches. [*character*]

lem

Lemmatized version of parliamentary speeches. Lemmatization was done using UDPipe analytical pipeline. [*character*]

codemp

Unique MPs code linking the corpus with the MPs meta-database which contains information on the background of most of the speakers. Missing codes are infrequent and primarily concern speakers who are not elected members of the Parliament. [*character*]

codeparty

Unique code assigned to a party which can be used to merge the corpus data with meta information collected on individual parties (e.g. party family or election result). [*character*]