Corpus of Debates in the Parliamentary Assembly of Bosnia and Herzegovina, 1998-2018

Michal Mochtak, Josip Glaurdić, Christophe Lesschaeve, and Ensar Muharemović

1 April 2022 (v1.0)

Contents

Overview

This is a codebook for the corpus of transcripts of parliamentary debates in the Parliamentary Assembly of Bosnia and Herzegovina [link]. Transcripts of speeches were text-mined and cleaned from machine-readable PDF documents using the R programming language. The corpus covers the period of 1998-2018 and counts six complete terms and over 127 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, Christophe Lesschaeve, and Ensar Muharemović (2022): BiHCorp: Corpus of Parliamentary Debates in Bosnia and Herzegovina (*vX.X*), https://doi.org/10.5281/zenodo.6517697.

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 Bosnia and Herzegovina.

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*]

house

House of Parliament; (DN) Dom naroda; (ZD) Zastupnički dom / Predstavnički dom. [*character*]

term1

Term duration. [*yyyy-yyyy*]

term2

Official numeric denominator for consecutive terms. The dataset covers six full terms (2nd - 7th). [*numeric*]

date

Date of a speech. If multiple days are recorded, the first day is used [date format: *yyyymmdd*]

date_raw

Date in its raw format. Multiple dates are preserved. [*character*]

meeting

Meeting descriptor [*character*]

agenda

Full description of agenda (if available). [*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 manually. 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*]

question

Dummy variable for the representatives' questions: (1) formal question; (0) others.

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. [*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 codes linking the corpus with the MPs meta database which contains information on 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/list which can be used to merge the corpus data with meta information collected on individual parties/electoral lists (e.g. party family or election result). [*character*]