Corpus of Debates in the Croatian Parliament, 2003-2020

Michal Mochtak, Josip Glaurdić, and Christophe Lesschaeve

1 April 2022 (v1.0)

Contents

Overview

This is a codebook for the corpus of transcripts of parliamentary debates in the Croatian Parliament [link]. The corpus covers the period of 2003-2020 and counts five complete terms and over 500 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): CROCorp: Corpus of Parliamentary Debates in Croatia ($vX.X$), https://doi.org/10.5281/zenodo.6521372.

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 Croatia.

Variables

The dataset contains 17 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 dataset covers five complete terms (5th - 9th). [*numeric*]

date

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

meeting

Number assigned to a meeting [*numeric*]

number

Number assigned to an agenda point. [*numeric*]

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 its application in the Croatian Parliament. The coding of agenda points was originally deployed by CEPIS [link]. As the codes covered only the first seven terms, we used them primarily as a comparative baseline in a paraphrase mining pipeline. The task was to extract potential labels for missing agenda points and manually check those that did not match. 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 [Culture]; (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 a speaker. *last name, first name* format. [*character*]

party

Party affiliation of a speaker. [*character*]

speech_link

Link to the transcript of a parliamentary speech. [*character*]

speech

Raw transcript of a parliamentary speech. [*character*]

lem

Lemmatized version of a parliamentary speech. Lemmatization was done using the 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*]