Proceedings of ParlaCLARIN III @ LREC2022 pages 131–134
Marseille, 20 June 2022

ParlaMint-RO: Chamber of the Eternal Future

Petru Rebeja[1], Mădălina Chitez[2] Roxana Rogobete[2], Andreea Dincă[2], Loredana Bercuci[2]
[1]Alexandru Ioan Cuza University of Iași, 11 Carol I Bvd., Iași, 700506
petru.rebeja@info.uaic.ro
[2] West University of Timisoara, 4 Vasile Pârvan St., Timișoara, 300223
madalina.chitez, roxana.rogobete, andreea.dinca, loredana.bercuci@e-uvt.ro

Abstract

The paper describes the ParlaMint-RO corpus of parliamentary debates in Romania. It analyses several trends in parliamentary debates (plenary sessions of the Lower House) held between 2000 and 2020. We offer a short description of the data collection, the workflow of data processing (text extraction, conversion, encoding, linguistic annotation), and an overview of the corpus. The paper then moves on to a multi-layered linguistic analysis, which offers an interdisciplinary perspective. We use computational methods and corpus linguistics approaches to scrutinize the future tense forms used by Romanian speakers in order to create a data-supported profile of the parliamentary group strategies and planning.

Keywords: ParlaMint-RO, linguistic analysis, future tense, data-supported parliamentary profile

1. Introduction

The discourse of the Romanian Parliament has been analysed by several studies, which have investigated such topics as the use of institutional forms of address (Ilie, 2010), epistemic markers (Ştefănescu, 2015), or situational argumentative strategies (Ionescu-Ruxăndoiu, 2015). These studies, however, do not conduct quantitative analyses, nor do they consult large corpora. Consequently, they do not offer a diachronic and statistical perspective. In this paper, we use a representative Romanian parliamentary discourse corpus, ParlaMint-RO for the first time. Compiled in the framework of the project ParlaMint - Towards Comparable Parliamentary Corpora, the corpus was financially supported by CLARIN ERIC[1], whose aim is to create free-access corpora of parliamentary discourse from as many as possible National Parliaments in Europe. ParlaMint corpora are created and encoded according to pre-established criteria and they are also uniformly encoded so that national datasets can be exchanged, reused and compared in different research scenarios (Erjavec et al., 2022).

The present study intends to introduce the ParlaMint Romanian sub-corpus (ParlaMint-RO) and to exemplify how the data can be used for interdisciplinary studies. After a short description of the data collection process, of the workflow of data processing, and an overview of the corpus, the paper will offer a linguistic analysis. We use computational methods to validate a study on the distribution of future tense forms across parliamentary groups. By analysing future tense forms, we aim to create a data-supported profile of some parliamentary groups' strategies and planning without extending the analysis towards the effectiveness of these strategies, i.e. whether future tense form use is associated with winning or losing parties (Kameswari and Mamidi, 2018). We extend, in this way, an exploratory study conducted by Grama (2022), which explored the discursive context of future tense forms in a corpus of interviews and press releases by Romanian local politicians. The study demonstrated that "promises for the better are made with every election season" (Grama, 2022, p. 31).

Level	Value
Number of transcribed sessions	1,832
Number of processed speeches	552,103
Number of words	109,304,196
Period	2000 – 2020

Table 1: Basic corpus statistics of ParlaMint-RO.

2. Data Collection: Parliamentary Records

The current corpus consists of transcripts of the plenary sessions of the Lower House, the Chamber of Deputies, as published on the official website[2]. Although both Romanian parliamentary chambers have published their transcripts (since 1996 for the Chamber of Deputies and since 2002 for the Senate), the structure of the source documents differs and is not very consistent. Therefore, as shown in Table 2, we limited the data selection to the Lower House. The time span covered ranges from 2000 to 2020 (five full legislative periods). The ParlaMint-RO corpus consists of 1833 files, one for each plenary session (a total of 1832 sessions, and the corpus root file), and comprises over 109 million words.

[1]Visit https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora for more information about the ParlaMint project.

[2]http://www.cdep.ro/pls/steno/steno2015.home

2.1. Processing Parliamentary Transcripts

The transcripts are published in HTML format and were received in bulk from the Information Technology Service of the Chamber of Deputies, after asking permission to use data for research and enquiring about additional / improved data sets. The data was extracted from the HTML files and converted to XML files using the `lxml` library[3] in Python. Alongside the default processing built into the lxml library (for encoding correction, HTML cleanup), we applied on-the-fly normalization of diacritics.

2.2. Corpus-Specific Metadata

The only metadata added to the corpus consists of the names, gender, and, for some deputies, their profile picture (if available). This data was scraped from the web site of the Lower House. The website also contains affiliation data. However, scraping that data resulted in a lot of erroneous records due to lack of a common structure in presentation. Thus, we did not include that data in the corpus since it needed to be manually corrected or supplemented by project members.

2.3. Encoding Transcripts into XML Format

The transcription began with an analysis of the ParlaMint schema, and in order for the team members to get accustomed to the schema, several sample sessions were manually coded by team members in Notepad+, according to the TEI format and documentation. The manual tagging made it possible to establish a set of tagging patterns and to extract specific recommendations for an automated process.

After identifying the patterns for locating and tagging sections of each transcription, we developed several Python scripts that automate the encoding of HTML transcriptions into XML format as much as possible:

- A crawler script downloads the names, gender and profile picture of the deputies,
- A parser script parses the session transcripts one by one and converts them into the XML format,
- Another script builds the corpus root file,
- After the corpus root file is built, another script is executed that applies the linguistic annotations to the existing corpus files, and creates the `.ana.xml` and `.conllu` files.

Despite our best efforts, we were not able to completely automate the encoding process. As such, after building the corpus root file, it still fails schema validation and needs manual intervention to correct the errors. Only after correcting the root file we could execute the script to perform linguistic annotation. Making the process fully automated is an ongoing task within the team.

The resulting XML files are structured according to the ParlaMint schema, which is based on the standard TEI structure[4], and is adapted to reflect the specific traits of Parliament sessions.

The source code for encoding raw transcripts from HTML format into the XML format required by the ParlaMint schema is available on Github[5], and will be updated to match the requirements of future versions and data.

2.4. Linguistic Annotation

The script that applies linguistic annotation iterates over the corpus files and queries the UDPipe Web API service[6] to perform tokenization, sentence segmentation, lemmatization, Part-of-Speech and morphological tagging, and dependency parsing. Unfortunately, the UDPipe service does not have a NER module for Romanian language so no NER was performed. We also tried to use the `spacy` library[7] which has a NER module available for Romanian but the library that converts the output from spacy to CoNNL-U format[8] has minor processing issues when used with the Romanian models. However, the results are not affected in the end.

3. Data Analysis

Since the future tense in Romanian is an analytical form, existing computational methods (such as UDPipe) extract the particular components of each form (auxiliary verb "to want" + root - infinitive form of the conjugated verb), therefore failing to automatically recognize the verbs we needed for the study. The difficulty of using UDPipe is represented by the fact that numerous verb tenses in Romanian are formed with auxiliary verbs, therefore the instrument cannot distinguish between different tenses built on the same auxiliary + root form, such as "will talk" ("voi vorbi" - formal future, indicative), "to talk" ("a vorbi" - infinitive), "talked" ("a vorbit" - compound perfect, indicative). The downfall of the low-resourced language as Romanian is that the data analysis requires more manual stages. As such, we decided to combine data from several sources for performing our analysis.

In the first step of data gathering, we downloaded the database dump from dexonline.ro[9], which contains Romanian word definitions that are not restricted by Intellectual Property rights. From the aforementioned database we extracted a list of 47,318 entries that were tagged as verbs.

For each of the extracted terms from the previous step, we try to obtain its inflections from conjugare.ro[10]. Retrieving data from conjugare.ro also validates whether the specified term is a verb or not; as such we narrowed

[3] https://lxml.de

[4] https://tei-c.org/

[5] https://github.com/romanian-parlamint/parsers

[6] http://lindat.mff.cuni.cz/services/udpipe/

[7] https://spacy.io/

[8] https://spacy.io/universe/project/spacy-conll

[9] Electronic version of Romanian Explanatory Dictionary, accessible at https://dexonline.ro

[10] htts://conjugare.ro

down the initial list of terms to 9,288 with more than 55,729 verb forms.

From the verb forms we selected the ones that represent the formal future tense (auxiliary + root, while omitting informal constructions such as particle "o" + conjunction "să" + conjugated verb in present: "o să vorbesc"; auxiliary verb "to have"/"a avea" + conjunction "să" + conjugated verb in present: "am să vorbesc"), this final list is then used to perform a cross-search on all the utterances from the corpora for the presence of each form. As such, we iterated through the whole corpus and built two sets of tuples from which we extracted the results: $(speaker, date, count_of_all_forms, count_of_all_words)$, and $(speaker, date, verb_form, count)$. Finally, we used `pandas`[11], and `matplotlib` libraries to aggregate the data and visualize the results.

The Python scripts, alongside the collection of verb forms are available on the Github page of our project[12].

4. Results

Romanian political discourse in general has been subject to various linguistic debates, mostly regarding the pragmatic or rhetoric dimension, such as stancetaking (Vasilescu, 2010), the practice of addressing (Saftoiu, 2013) or even verbal aggressiveness (Roibu and Constantinescu, 2010). In contrast, our data analysis focuses on a more specific topic - the distribution of future tense forms. It seems that a common rhetorical strategy in the Romanian Parliament is to refer to future projects or broader aims rather than ongoing projects. This permanent projection is not a sign of activism or concern for future policies. In different contexts, studies (Bertrand, 2021) have shown it is a sign of non-engagement, of the lack of solid commitment and of a tendency to delay actions. Unlike other languages, the cases when the Romanian present tense marks prospective actions are to be found mostly in literary texts - thus in stylistically rich contexts.

4.1. Verb Analysis: Future Tenses

Our analysis revealed the identity of the 10 politicians who use future tenses most frequently (4.1).

The 10 speakers, some of whom have shifted allegiances, were at the time of their speeches affiliated with the following parties: PSD, PNL, PRM (4.1).

Six of the speakers are affiliated with the Social Democratic Party (PSD), the largest in the country, which held the majority and control in most of the legislatures. This explains their high number of interventions (and the total number of future tense verbs: 33,728). The party's discourse consists of verbs of action projected into the future. Another four speakers are members of the National Liberal Party (PNL), with a total of 11,169 future verbs. One speaker belongs to the far-right nationalist party, Great Romania Party (PRM), which was not present in all national mandates.

Examples of use show a lack of tangible projects for the development of the country: "We will never again guarantee the governmental assumption of responsibility"; "Let's all think about the many and we'll see that we really are a different kind of politicians.". Moreover, when analysing the most frequent nouns and verbs present in the corpus, we noticed a preference for terms usually present in law voting procedure and meeting agenda ("law", "committee"), discourse mark-

Speaker	Count	Pct
Valer Dorneanu	10,859	0.73
Tudor Ciuhodaru	7,842	1.55
Emil Boc	4,480	1.48
Valeriu Ștefan Zgonea	4,190	0.51
Florin Iordache	3,837	0.62
Adrian Moisoiu	3,775	0.98
Doru Ioan Tărăcilă	3,602	0.73
Gheorghe-Eugen Nicolăescu	3,421	1.40
Nicolae Văcăroiu	3,398	0.85
Bogdan Olteanu	3,268	0.76

Table 2: Most frequent users of future tenses. The column *Count* displays the total number of future forms used by a speaker, and the column *Pct* shows the percentage of future forms from the total number of words spoken by the same person.

Speaker	Affiliation and time-span
V. Dorneanu	PDSR/PSD-Social Democratic Party (2000–2008)
T. Ciuhodaru	PSD/Independent/ PPDD-People's Party–Dan Diaconescu (2008–2016)
E. Boc	PD-Democratic Party – now PNL-National Liberal Party (2000–2004)
V. Ș. Zgonea	PSD-Social Democratic Party (2000–2016)
Fl. Iordache	PDSR/PSD-Social Democratic Party (2000–2020)
A. Moisoiu	PRM-Greater Romania Party (2000–2008)
D. I. Tărăcilă	PSD-Social Democratic Party (2000–2008)
Gh.-E. Nicolăescu	PNL-National Liberal Party (2000–2017)
N. Văcăroiu	PSD-Social Democratic Party (2000–2008)
B. Olteanu	PNL-National Liberal Party (2004–2009)

Table 3: Affiliation of the 10 politicians who most frequently use future tenses.

[11] https://pandas.pydata.org/

[12] https://github.com/romanian-parlamint/future-tense-usage

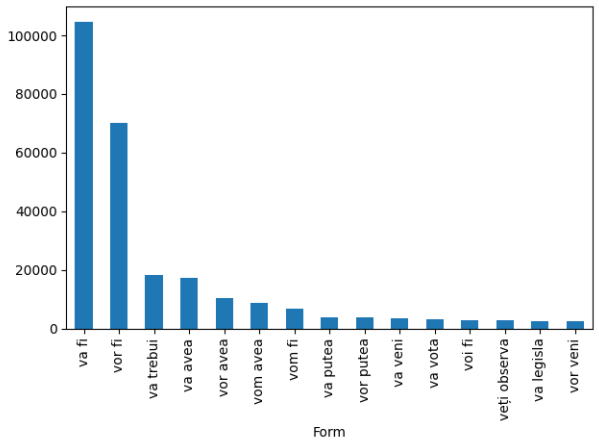

Figure 1: Top 15 inflections in future tense used in Lower House transcriptions.

ers ("thank" as a closing remark, direct addresses such as "mister president"), but likewise general ones that focus on the state of the country: "Romania", "project", "state", "years" etc. The top 15 inflections in future tense (1) reveals only three forms in first person ("will have", "will be" - singular and plural) and eleven in third person (either singular or plural: "will be", "will have to", "will be able to", "will come", "will vote", "will legislate"), which suggests an impersonal tone related to shifting responsibility onto others. Another sign of projection apparent in the deputies' speeches is the frequent use of "ar trebui să", a conditional tense that can be translated with "should".

5. Conclusions and Future Work

The present corpus still needs adjustments in order to obtain accurate data and optimize the workflow. Additionally, the linguistic analysis should be expanded and detailed in future studies in order to make more verb patterns available. We also had several difficulties in processing such large amounts of data with corpus linguistics tools that do not involve programming skills. When the ParlaMint-RO corpus is completed, numerous research directions can be pursued, such as investigating direct addressing, appellations used during debates (divided by parties and gender), or parliamentary topics and political ideology, thus opening valuable pathways for comparative research in political, linguistic or intercultural studies.

6. Acknowledgements

The study would have not been completed without the data provided by online resources such as dexonline.ro and conjugare.ro. The ParlaMint-RO corpus was compiled in the framework of the ParlaMint project supported by CLARIN ERIC.

7. Bibliographical References

Bertrand, D. (2021). Future or past future tense? what political timeframe? *E—C*, (32):34–41, Nov. https://mimesisjournals.com/ojs/index.php/ec/article/view/1500.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., et al. (2022). The parlamint corpora of parliamentary proceedings. *Language resources and evaluation*, pages 1–34. https://doi.org/10.1007/s10579-021-09574-0.

Grama, E.-M. (2022). The language of romanian administration: an interview-based corpus case study. In Madalina Chitez, et al., editors, *Corpus Related Digital Humanities: Interdisciplinary Micro Perspectives*, pages 27–32, Timișoara. Editura Universității de Vest.

Ilie, C. (2010). Managing dissent and interpersonal relations in the romanian parliamentary discourse. *European parliaments under scrutiny*, pages 193–223. https://doi.org/10.1075/dapsac.38.11ili.

Ionescu-Ruxăndoiu, L. (2015). Discursive perspective and argumentation in the romanian parliamentary discourse. a case study. *L'Analisi Linguistica e Letteraria 2008-1*, 16:435–441. https://www.analisilinguisticaeletteraria.eu/index.php/ojs/article/view/423/359.

Kameswari, L. and Mamidi, R. (2018). Political discourse analysis: A case study of 2014 andhra pradesh state assembly election of interpersonal speech choices. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. https://aclanthology.org/Y18-1028.pdf.

Roibu, M. and Constantinescu, M. N. (2010). Verbal aggressiveness in the romanian parliamentary debate. past and present. *Revue roumaine de linguistique*, 54(4):353–364. https://www.lingv.ro/RRL%204%202010%20art04Roibu.pdf.

Saftoiu, R. (2013). The discursive practice of addressing in the romanian parliament. *The Pragmatics of Political Discourse: Explorations across Cultures. Amsterdam: John Benjamins Publishing*, pages 47–68. https://doi.org/10.1075/pbns.228.04saf.

Ştefănescu, A. (2015). Analysing the rhetoric use of the epistemic marker eu cred că (i think) in romanian parliamentary discourse. *Persuasive Games in Political and Professional Dialogue*, 26:101. https://doi.org/10.1075/ds.26.06ste.

Vasilescu, A. (2010). Metastance in the romanian parliamentary discourse: Case studies. *The proceeding of Institutul de Lingvistica al Academiei. LV*, 4:365–380. https://www.lingv.ro/RRL%204%202010%20art05Vasilescu.pdf.