ParlaMint-RO: Chamber of the Eternal Future

Petru Rebeja[1]   Mădălina Chitez[2]   Roxana Rogobete[2]
Andreea Dincă[2]   Loredana Bercuci[2]

Language Resources and Evaluation Conference
June 20, 2022

[1]Alexandru Ioan Cuza University of Iași

[2]West University of Timișoara

Introduction

- Study the distribution of future tense forms across parliamentary groups,

- Perform a quantitative analysis of large corpus,

- Exemplify how the data can be used for interdisciplinary studies.

Data Collection: Parliamentary Records

Processing Transcriptions

- Transcripts from HTML format are converted to XML format required by ParlaMint schema using the `lxml` library[1].

- On-the-fly normalization of text is applied to extracted data.

---

[1]https://lxml.de

Corpus-Specific Metadata

- Names, gender, and picture of deputies were scraped from the website of the Lower House[2]

- Scraping affiliation data was unsuccessful; as such, no affiliation data is included in the corpus.

[2]http://www.cdep.ro/

Basic statistics of ParlaMint-RO corpus

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

Data Analysis

Data Sources of Future Tense Usage Analysis

- ParlaMint-RO transcriptions,
- List of verbs from `https://dexonline.ro`[3],
- Inflected forms of verbs for *formal future tense* from `https://conjugare.ro`.

---

[3]The list is restricted only to dictionary entries that are not under Intelectual Property rights.

Data Normalization

The data sets were joined and normalized into two sets of tuples:

(*speaker*, *date*, *count of all forms*, *count of all words*)

(*speaker*, *date*, *verb form*, *count*)

Data Aggregation

The tuples were aggregated using *pandas*[4], and *matplotlib*[5] was used for visualization.

The Python scripts, alongside the collection of verb forms are available on Github[6].
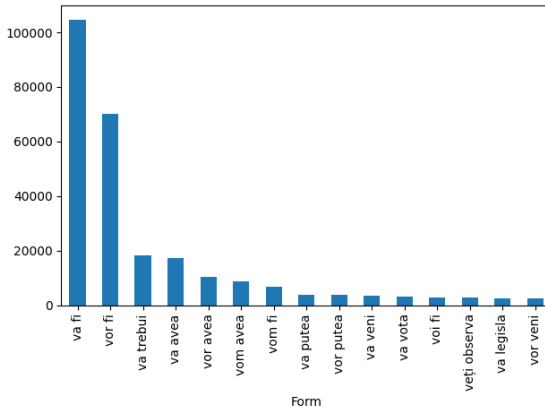
---

[4] https://pandas.pydata.org
[5] https://matplotlib.org
[6] https://github.com/romanian-parlamint/future-tense-usage

Results

Top 10 Most Frequent Users of Future Tense Forms

Speaker	Count	Percentage
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

Top 15 Inflections In Future Tense Used In Lower House

va fi	(he/she) will be	vom avea	(we) will have	va vota	(he/she) will vote
vor fi	(they) will be	vom fi	(we) will be	voi fi	(I) will be
va trebui	will have to	va putea	(he/she) will be able to	veți observa	(you) will observe
va avea	(he/she) will have	vor putea	(they) will be able to	va legisla	(he/she) will legislate
vor avea	(they) will have	va veni	(he/she) will come	vor veni	(they) will come

- The top 15 inflections in future tense reveals only three forms in first person, 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".

Q&A

Thank you!