Look Who's Talking: The Most Frequently Used Words in the Bulgarian Parliament 1990-2024

Abstract

This study identifies the most frequently used words and some multi-word expressions in the Bulgarian Parliament. We do this by using the transcripts of all plenary sessions between 1990 and 2024 - 3,936 in total. This allows us both to study an interesting period known in the Bulgarian linguistic space as the years of "transition and democracy", and to provide scholars of Bulgarian politics with a purposefully generated list of additional stop words that they can use for future analysis. Because our list of words was generated from the data, there is no preconceived theory. We include all interactions during all sessions, our analysis goes beyond traditional party lines. We provide details of how we selected, retrieved, and cleaned our data, and discuss our findings.

Ruslana Margova

GATE Institute,

Sofia University St. Kliment Ohridski

ruslana.margova@gate-ai.eu

Bastiaan Bruinsma

Chalmers University

of Technology

sebastianus.bruinsma@chalmers.se

Keywords: ***corpus, parliament, most frequently used words, Bulgaria***

1. Object and motivation

- The political changes in Bulgaria in 1989 led to greater transparency of political power, including the right of public access to information=> transcripts of National Assembly were made public.
- Natural language processing (NLP) methods to study this corpus as a whole for:
- better understanding of the topics discussed by parliamentarians
- salience theory (Budge and Farlie, 1977), frequently used words are of greater importance to speakers and can provide insight into the interests of the National Assembly.

2. Background

- In the Bulgarian context, the focus is often on individual political speeches; studies on the use of foreign words; the media behaviour of the political elite; linguistic aggression; the appearance of European identity; the use of clichés, dialects and factual errors; the quantitative ratio of words from one national assembly to another; the language of certain MPs.
- Various attempts to expand the current corpus: an annotated version of part of the corpus; use the audio of the speeches to build a new corpus of Bulgarian speech suitable for training and evaluating modern speech recognition systems; ParlaMint dataset

3. Data and Pre-processing

- Each of the 3,936 minutes is structured in the same way.
- After downloading transcripts were convert from HTML to TXT format; remove headers and footers; tokenise our words (this and all subsequent steps are performed using version 3.3.1 of the quanteda package in R; lowercase them, generate n-grams to capture common expressions, remove punctuation, symbols and numbers, and finally remove stop words as contained in the BulTreeBank corpus.
- Result: a corpus of 694,174 unique tokens and we focus on the 250 most frequent words in this resulting data set (the last of which had a relative frequency of 0.033%), although this cut-off is necessarily arbitrary.

4. Results:

a) legal terms;

The most common type (in terms of frequency) are words related to law, where the two abbreviations "ал'' (paragraph) and "чл'' (article) are the most common, followed by "закон'' (law), "запонопроект'' (draft) and "предложение'' (proposal).

b) places and countries;

Unsurprisingly, the word "България'' (Bulgaria) is the most frequently used, followed by related terms such as "страна'' (country), "държава'' (state), "република'' (republic), "българските'' (Bulgarian - adjectival) and "граждани'' (citizen). Bulgarian as a nationality does not appear in this list of most frequently used words, but can be found instead in references to "общество'' (society) or "хора'' (people). More geographical references - such as "Европейският съюз'' (European Union) and "София'' (Sofia) - can also be found. It is noteworthy that Osenova2012a found similar terms, suggesting that these terms have changed little in importance over time.

c) financial;

Another common category is financial references - most often to the Bulgarian currency ("лв''). We also find words such as "пари'' (money), "бюджет'' (budget), "хиляди'' (thousands) and "милиони'' (million). Note that there are no references to other currencies. This suggests that the debate on the adoption of the euro as the official currency is not (yet) dominant during the period we are studying.

d) parliamentary behaviour;

Next, we find words that demonstrate politeness and respect for colleagues (see Osenova 2012a, Tarasheva2015a), where we find words such as "уважаеми'' (dear), "моля'' (please), and "благодаря'' (thank you). This kind of politeness is often nothing more than a set of linguistic conventions that operate independently of the current goal a speaker is trying to achieve (Christie2002a). As such, this type of politeness is more operational, helping politicians to introduce themselves, rather than reflecting their opinions of each other.

e) procedural;

Related to this are words that refer to different parliamentary procedures, such as "решение'' (decision), "гласуване'' (voting), "комисия'' (commission), "изказвания'' (speeches), "предложения'' (suggestions), "въпрос'' (question), "процедура'' (procedure), реплики (replies), and "текстове'' (texts).

f) verbs;

We find words like "мисля'' (think), "казвам'' (say), "смятам'' (consider), "разбира'' (understand), and "искам'' (want)

g) adverbs;

Such as "всъщност'' (in fact), "наистина'' (really), "ясно'' (clearly), "просто'' (simply), "тоест'' (i.e.), "действително'' (actually), "изключително'' (exceptionally), and "вярно'' (truly). Interestingly, there are no verbs expressing insistence. Instead, the imperative particle "нека'' (let us) is often used. Moreover, the tendency to use impersonal constructions also shows that parliamentarians seem to be trying to avoid personal responsibility, opting instead for general responsibility.

h) party abbreviations;

Finally, we find references to the parties. Interestingly, although the corpus consists of texts from more than 30 years, the word ГЕРБ - an abbreviation of one of the political parties - is also among the most frequently used words ("Граждани за европейско развитие на България'' - Citizens for European Development of Bulgaria).

j) other.

4. Conclusions and future work

- Bulgarian politicians use Bulgaria prominently in their speeches;
- terms such as "European'' are also important, but not as central as "Bulgarian'';
- the speeches also show linguistic politeness, presumably as a convention.
- abbreviations related to law are common, as are terms describing procedures in legislative tasks
- verbs indicating cognitive effort are widespread, but the frequent use of the imperative particle "нека'' (let us) suggests a tendency to defer decision-making or responsibility
- the abbreviation for the Bulgarian currency is noteworthy, while the dominance of the abbreviation for the political party "ГЕРБ'' reflects the dominance of this particular party, despite the presence of others in Parliament during the period analysed.
- The generated list contains meaningful words such as "budget'', "decision'', "abstention'', "understand'', which are semantically relevant and essential and cannot be considered as stop words.

The list can be used for specific purposes for further automated linguistic analysis with a different focus: for in-depth analysis of the main themes in the contemporary development of politics and public attitudes in Bulgaria after the beginning of the democratic changes; how language has changed over the years; comparative analysis of the language of individual parties on particular issues.

Acknowledgements: The results presented in this paper are part of the GATE Project. This project has received funding from the European Union's Horizon 2020 WIDESPREAD- 2018-2020 TEAMING Phase 2 programme under Grant Agreement No. 857155, and partly by BROD project 101083730.