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

Ruslana Margova
GATE Institute
Sofia University "St. Climent Ohridsky"
ruslana.margova@gate-ai.eu

Bastiaan Bruinsma
Chalmers
University of Technology
sebastianus.bruinsma@chalmers.se

Abstract

In this study we identify 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, and because 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.

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

1 Object and motivation

The political changes in Bulgaria in 1989 led to demands for greater transparency of political power, including the right of public access to information. As a result, transcripts of meetings of the National Assembly ([1]) were made public, but only after considerable public pressure. So far, these transcripts have mainly been used for qualitative analysis of individual debates on a case-by-case basis. They have rarely been considered as a corpus in their own right, most likely due to the considerable number of transcripts available (every transcript from 27 February 1879 - just 17 days after the National Assembly was established - and onwards is available) as well as the way they were made accessible (each had to be downloaded individually).

Here, we will use natural language processing (NLP) methods to study this corpus as a whole, allowing us to identify the most frequently used words in the National Assembly between 1990 and 2024. We do this for both theoretical and methodological reasons. The theoretical reasons include gaining a better understanding of the topics discussed by parliamentarians. According to 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. This is particularly relevant as the period under study witnessed significant and structural changes in Bulgarian politics and society. Methodological reasons include our desire to generate a list of stop words that future researchers can use to further preprocess this corpus to better estimate any concepts of interest, as well as to provide an example of how this data can be used for other NLP-related problems.

2 Background

Transcripts of legislative debates are often used to study the opinions, positions and policy preferences of elected politicians (Abercrombie and Batista-Navarro, 2020). In the Bulgarian context, the focus is often on individual political speeches and the debate in which they were made. Thus, studies have been conducted on the use of foreign words (Rachev, 2023), the media behaviour of the political elite (Todorov, 2001; Yurukova, 2022), linguistic aggression (Uzanicheva, 2020; Milanov, 2021; Nenova, 2021), the appearance of European identity (Mavrodieva, 2014), the use of clichés, dialects

[1] https://www.parliament.bg/bg/plenaryst

and factual errors (Milanov and Mihailova-Stalyanova, 2022), the quantitative ratio of words from one national assembly to another (Tarasheva, 2017), and the language of certain MPs (Tarasheva, 2015).

In addition, various attempts have been made to expand the current corpus. For example, (Osenova and Simov, 2012) provide an annotated version of part of the corpus; (Geneva et al., 2019) use the audio of the speeches to build a new corpus of Bulgarian speech suitable for training and evaluating modern speech recognition systems; and the Strazha Foundation will combine it with the duration of each session, the number of words by party, the average number of words per MP by party, the most verbose MPs and other related facts to discuss and comment on the current state of the National Assembly ([2]).

Finally, transcripts from 2015 onwards have been made part of the ParlaMint dataset (Erjavec et al., 2023), in which each political speech is annotated with, among other things, the age, gender and political orientation of the speakers. As ParlaMint contains similar data from 17 European national parliaments, this allows for cross-country comparisons, as shown by Miok et al. (2023).

One thing these transcripts have not been used for is to examine the frequency of word choice. This is interesting, as this is often seen as one of the basic requirements for understanding the corpus (O'Keeffe and McCarthy, 2010). As a result, a domain-specific list of words that can be used as stop words is missing, as this requires recourse to the corpus one wishes to use (Sarica and Luo, 2021; Yang and Wilbur, 1996). Thus, the creation of such a list can help scholars to better deal with the data from these transcripts and make future analyses less complicated.

3 Data and Pre-processing

Each of the 3,936 minutes is structured in the same way. First, the chair and vice-chair and the secretary are identified, together with the date and time of the meeting. Then each speaker is identified individually and their remarks are listed. This includes both what they say and what else is happening in the meeting at the same time. However, while noise or applause is included with general remarks, the specific insults and attacks from the floor are not (Tarasheva, 2017: cf.). The transcripts do not record the insults exchanged by the deputies in the chamber, but only those uttered from the gallery. The meetings themselves have no particular structure - sometimes votes are followed by further discussion; sometimes meetings begin with an agenda, but not always; sometimes they begin with proposals to change the agenda; and sometimes there are agenda items listed at the beginning.

After downloading the individual transcripts from the National Assembly website ([3]), we convert them from HTML to TXT format. We fix any encoding problems and remove headers and footers. Next, we tokenise our words (this and all subsequent steps are performed using version 3.3.1 of the quanteda package in R (Benoit et al., 2018)), 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 (Simov, 2014). This last step is crucial, as failure to do so would result in the identification of stop words that are common to Bulgarian in general, rather than those that are specific to the National Assembly. It also prevents our multi-word expressions (MWEs) from consisting solely of collections of frequently used words and expressions. This results in a corpus of 694,174 unique tokens. For our purposes here, 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. Appendix A provides an overview of these words, together with an English translation.

4 Results

As a result, we get a list of words and some typical MWEs for parliamentary speeches. There is no specific study of MWEs in this analysis. However, MWEs and their derivatives play an important role in certain topics when NLP methods are used (Barbu Mititelu and Leseva, 2018). The list is rich with collocations typical for parliamentary life such as ”уважаеми

[2] https://www.strazha.bg/

[3] https://www.parliament.bg/bg/plenaryst

дами господа народни представители" (Respectfully, ladies and gentlemen deputies), "предложението прието" (The proposal is accepted).

Through a political-historical prism we can distinguish nine groups of meaning-functional types of words in the resulting list: a) legal terms; b) places and countries; c) financial; d) parliamentary behaviour; e) procedural; f) verbs; g) adverbs; h) party abbreviations; j) other. These types are not surprising. In an earlier study on the Bulgarian language in general, Koeva et al. (2012) found that the most commonly used nouns are those related to time, place and people.

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). This is followed by geographical references. 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 Osenova and Simov (2012) found similar terms, suggesting that these terms have changed little in importance over time. 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.

Next, we find words that demonstrate politeness and respect for colleagues (Osenova and Simov, 2012; Tarasheva, 2015: see also), 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 (Christie, 2002). As such, this type of politeness is more operational, helping politicians to introduce themselves, rather than reflecting their opinions of each other. 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).

Two other categories are verbs and adverbs. Under the former, we find words like "мисля" (think), "казвам" (say), "смятам" (consider), "разбира" (understand), and "искам" (want), and under the latter words 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.

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). And while the word "герб" can also refer to a coat of arms, in the parliamentary context here there is no doubt that the disambiguation of the word refers to the political party.

5 Conclusions and future work

The analysis of a corpus of Bulgarian parliamentary speeches reveals some interesting findings: 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. However, the additional list provided can be used for specific purposes for further automated linguistic analysis with a different focus: for example, for more 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. The large dataset allows for the study of how language has changed over the years, as well as for comparative analysis of the language of individual parties on particular issues. A more in-depth study can reveal the MWEs in parliamentary speech and their pragmatic role.

6 Acknowledgments

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.

References

Gavin Abercrombie and Riza Batista-Navarro. 2020. Sentiment and Position-Taking Analysis of Parliamentary Debates: A Systematic Literature Review. *Journal of Computational Social Science*, 3:245–270.

Verginica Barbu Mititelu and Svetlozara Leseva. 2018. Derivation in the Domain of Multiword Expressions. In Manfred Sailer and Stella Markantonatou, editors, *Multiword Expressions Insights from a Multi-Lingual Perspective*, pages 215–246. Language Science Press, Berlin.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An R Package for the Quantitative Analysis of Textual Data. *Journal of Open Source Software*, 3(30):774.

Ian Budge and Dennis Farlie. 1977. *Voting and Party Competition*. John Wiley & Sons, London.

Chris Christie. 2002. Politeness and the Linguistic Construction of Gender in Parliament: An Analysis of Transgressions and Apology Behaviour. Sheffield Hallam Working Papers on the Web: Linguistic Politeness and Context.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkadhur Barkarson, Steinthór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2023. The ParlaMint Corpora of Parliamentary Proceedings. *Language Resources and Evaluation*, 57(1):415–448.

Diana Geneva, Georgi Shopov, and Stoyan Mihov. 2019. Building an ASR Corpus Based on Bulgarian Parliament Speeches. In *Statistical Language and Speech Processing: 7th International Conference, SLSP 2019, Ljubljana, Slovenia, October 14–16, 2019, Proceedings*, pages 188–197, Berlin, Heidelberg. Springer-Verlag.

Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova, and Ekaterina Tarpomanova. 2012. The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, 0(1):65–110.

Ivanka Mavrodieva. 2014. The Concept of the Bulgarian "European Identity" Across the Bulgarian Language and Political Rhetoric in Bulgaria. *US-China Foreign Language*, 12(1):1–16.

Vladislav Milanov. 2021. Българската политическа реч в балкански и славянски контекст. *Studia Philologica*, 40(1):265–278.

Vladislav Milanov and Nadezhda Mihailova-Stalyanova. 2022. *Езикови портрети на български политици [Linguistic Portraits of Bulgarian Politicians]*. "St. Kliment Ohridski" University Publishing House, Sofia.

Kristian Miok, Encarnacion Hidalgo-Tenorio, Petya Osenova, Miguel-Angel Benitez-Castro, and Marko Robnik-Sikonja. 2023. Multi-aspect Multilingual and Cross-lingual Parliamentary Speech Analysis.

Kristina Nenova. 2021. Езикът на омразата – изследване на политическата риторика и отражението й върху качеството в медийната среда. In *Качествена журналистика и нова комуникационна среда [Quality journalism and a new communication environment]*, pages 295–304, Sofia. Faculty of Journalism and Mass Communication, Sofia University "St. Cl. Ohridski".

Anne O'Keeffe and Michael McCarthy. 2010. *The Routledge Handbook of Corpus Linguistics*. Routledge, London.

Petya Osenova and Kiril Simov. 2012. The Political Speech Corpus of Bulgarian. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1744–1747, Istanbul, Turkey. European Language Resources Association (ELRA).

Yordan Rachev. 2023. Чуждиците На Политиците. In *2023: Език, наука, комуникации и спорт – 60 години академично образование. Сборник с доклади [2023: Language, Science, Communications and Sport - 60 years of academic education. Collection of reports]*, pages 158–164, Varna. Varna Medical University Press.

Serhad Sarica and Jianxi Luo. 2021. Stopwords in Technical Language Processing. *PLOS ONE*, 16(8):1–13.

Kiril Simov. 2014. BulTreeBank Stopword List. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Elena Tarasheva. 2015. Изследване на дискурса на Волен Сидеров в 42. Народно събрание. *Rhetoric and Communications*, October(19):1–10.

Elena Tarasheva. 2017. 43 Народно събрание: Двойно повече приказки, Нинова говори пред изключен микрофон, Борисов спира руски проекти. Accessed: 12-04-2024.

Svetoslav Todorov. 2001. Медийната изява на лидерите на политически партии в България. *Социологически проблеми*, 33(3+4):137–153.

Yuliana Uzanicheva. 2020. Стратегии на езикова агресия в парламентарната реч на България и Украйна. *Българска реч. Списание за езикознание и езикова култура*, 1:42–49.

Yiming Yang and John Wilbur. 1996. Using Corpus Statistics to Remove Redundant Words in Text Categorization. *Journal of the American Society for Information Science*, 47(5):357–369.

Maria Yurukova. 2022. *Изборите За Ен През 2019 Г. Отразяване В Българските Онлайн Медии [EP Elections In 2019 Coverage In The Bulgarian Online Media]*. "St. Kliment Ohridski" University Publishing House, Sofia.

Appendix A Word List

Term	Translation
ал	paragraph
чл	article
българия	Bulgaria
заповядайте	please
уважаеми колеги	dear colleagues
лв	BNG
против	against
уважаеми господин председател	dear_mr_president
отношение	attitude
закон	law
господин председател	mr_president
става	happen
мисля	think
колеги	colleagues
наистина	truly
начин	a way
именно	namely
разбира	of course
въпрос	question
хора	people
хората	the people
предложението прието	the proposal is accepted
тоест	i.e.
комисията	the commission
държавата	the country
страна	the country
имаме	we have
включително	included
част	part
връзка	connection
закона	the law
изказвания	statements
народното събрание	parliament
просто	simply
предложение	suggestion
уважаема госпожо председател	dear Mrs president
господин министър	dear minister
текст	text
страната	the country
знаете	you know
всъщност	in fact

continued on next page

continued from previous page

Term	Translation
смятам	I believe
кажа	I say
решение	decision
реплики	replica
правителството	government
комисията подкрепя предложението	the commission supports the proposal
законопроект	draft bill
ясно	clear, obvious
относно	regarding
виждам	i see
свързани	linked, connected
гласувайте	please, vote!
средства	meanings
госпожо председател	MRS president
път	way
предложението	suggestion
следва	then
нека	let
процедура	procedure
залата	the hall
въпросът	the question
стане	it will happen
говорим	we are talking about
неща	things
народни представители	MP
времето	the time
право	law
имате думата	you have the floor
казвам	I say
информация	information
означава	it means
пари	money
съответно	thus
предложения	suggestion
господин	Mr
лица	faces
практика	practice
гласуваме	we are voting
работа	work
предлага	suggestion
въпроси	quesions
уважаеми дами господа народни представители	dear MP

continued on next page

continued from previous page

Term	Translation
законопроекта	draft bill
проблем	problems
казва	i say
става дума	it means
възможност	possibility
млн	millions
място	place
знам	i know
думата	word
въздържали	abstention in voting
въпроса	question
действително	really
комисията подкрепя текста вносителя	the commission supports the proposal
дейност	activity
заменят	change
човек	person
друго	other
народните представители	MP
такава	such
рамките	frame
член	article
съжаление	regret
уважаеми народни представители	dear MP
комисия	commission
случай	case
проект	project
хил	thousand
работи	work
имаше	had
необходимо	necessary
надявам	hope
говори	speak
бюджета	budget
второ	second
момент	moment
става въпрос	it means
предлагам	i suggest
реплика	replica
правим	we make
европейския съюз	European Union
уважаеми господин министър	dear MP
текста	text
парламента	the parliament
министерския съвет	council of ministers
промени	change

continued on next page

continued from previous page

Term	Translation
искате	want
вносител	importer
цел	target
можем	we can
правят	they do
проблеми	problems
изключително	exceptional
данни	data
резултат	result
министър	minister
текстове	text
смисъл	meaning
достатъчно	enough
определени	particular
такова	such
трябвало	should
политика	politics
срок	deadline
искам	want
общините	municipalities
случаи	cases
законът	law
иначе	otherwise
очевидно	obvious
против въздържали	against
приема	accept
колегите	colleagues
система	system
вниманието	attention
зала	hall
управление	government
думите	word
мерки	measure
общо	general
независимо	independent
гласуване	voting
работата	work
дейности	activities
предложението приема	suggestion
същото	same
контрол	control
софия	capital
направим	we make
процедурата	procedure
ред	order
възможността	possibilities
принцип	principal

continued on next page

continued from previous page

Term	Translation
дейността	activity
извършва	making
промяна	change
вчера	yesterday
република българия	Bulgaria
абсолютно	total
герб	GERB
какви	which
казах	said
случая	case
каже	say
значи	means
решения	decision
оглед	meaning
бюджет	budget
българските граждани	Bulgarians
нещата	things
случи	happened
другото	others
създава	creates
държава	country
отсъства	are missing
различни	different
условия	cases
лицата	faces
другите	others
решението	decision
имате	you have
документи	documents
единствено	only
страни	different
едни	ones
т.н	etc
последните	last
програма	program
струва	costs
работят	work
правото	law
искаме	want
членове	participants
своите	their
разходи	costs
б	b
искам кажа	want to say
дава	gives
цели	goals
положение	position

continued on next page

continued from previous page

Term	Translation
лично	personal
системата	system
обществото	society
доклада	report
предвижда	foresee
средствата	means
действия	works
фонд	fund
казахте	said
началото	at the beginning
съгласно	according to
подкрепа	supported
тема	topic
нататък	follow
крайна сметка	at the end
приет	accepted
политически	political
някакви	some
води	leads
гражданите	citizens
възможно	possible
господин_димитров	Mr. Dimitrov
вярно	really
трябваше	it should be
процес	processes
договор	contract
съответните	respectively
отговор	answer

Table 1: Overview of the 250 most frequent words, their frequency and translation