1.01 UDC: 003.295: 342.537.6(497.4)"2014/2018"

Darja Fišer,[*] Nikola Ljubešić,[**] Tomaž Erjavec[***]

Parlameter – a Corpus of Contemporary Slovene Parliamentary Proceedings

IZVLEČEK

PARLAMETER – KORPUS RAZPRAV SLOVENSKEGA DRŽAVNEGA ZBORA

V prispevku predstavimo korpus sodobnih parlamentarnih razprav Parlameter, ki vsebuje razprave 7. mandata slovenskega Državnega zbora (2014–2018). Korpus Parlameter vsebuje bogate metapodatke o govorcih (spol, starost, izobrazba, strankarska pripadnost) in je jezikoslovno označen (lematizacija, tegiranje), kar omogoča številne raziskave s področja digitalne humanistike in družboslovja. V prispevku prikažemo potencial korpusnoanalitičnih tehnik za raziskovanje političnih razprav. Korpusna arhitektura je zasnovana tako, da omogoča širitev korpusa na druga časovna obdobja, prav tako pa tudi vključevanje gradiv drugih parlamentov, začenši s hrvaškim in bosanskim.

Ključne besede: parlamentarne razprave, izdelava korpusa, jezikovne tehnologije, korpusna analiza

ABSTRACT

The paper presents the Parlameter corpus of contemporary Slovene parliamentary proceedings, which covers the VIIth mandate of the Slovene Parliament (2014–2018). The Parlameter corpus offers rich speaker metadata (gender, age, education, party affiliation)

* **Department of Translation, Faculty of Arts, University of Ljubljana, Aškerčeva cesta 2, SI-1000 Ljubljana, Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, darja.fiser@ff.uni-lj.si**

** **Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, nikola.ljubesic@ijs.si**

*** **Department of Knowledge Technologies, Jožef Stefan Institute, Jamova Cesta 39, SI-1000 Ljubljana, tomaz.erjavec@ijs.si**

and is linguistically annotated (lemmatization, tagging), which boost research in several digital humanities and social sciences disciplines. We demonstrate the potential of the corpus analysis techniques for investigating political debates. The corpus architecture allows for regular extensions of the corpus with additional Slovene data, as well as data from other parliaments, starting with Croatian and Bosnian.

Keywords: parliamentary proceedings, corpus construction, language technology, corpus analysis

Introduction

Parliamentary discourse is motivated by a wide range of communicative goals, from position-claiming, persuasion and negotiation to agenda-setting and opinion-building along ideological or party lines. It is characterized by role-based commitments and confrontation and the awareness of a multi-layered audience (Ilie 2017). The unique content, structure and language of records of parliamentary debates are all factors that make them an important object of study in a wide range disciplines in digital humanities and social sciences, such as political science (van Dijk 2010), sociology (Cheng 2015), history (Pančur and Šorn 2016), discourse analysis (Hirst et al. 2014), sociolinguistics (Rheault et al. 2016), and multilinguality (Bayley 2014).

Despite the fact that parliamentary discourse has become an increasingly important research topic in various fields of digital humanities and social sciences in the past 50 years (Chester and Bowring 1962; Franklin and Norton 1993), it has only recently started to acquire a truly interdisciplinary scope (Bayley 2014). Recent developments enable cross-fertilization of linguistic studies with other disciplines and in-depth exploration of institutional uses of language, interpersonal behaviour patterns, interplay between language-shaped facts, and reality-prompted language ritualization and change (Ihalainen et al. 2016).

With an increasingly decisive role of parliaments and their rapidly changing relations with the public, mass media, executive branch and international organizations, further empirical research and development of integrative analytical tools are necessary in order to achieve a better understanding of parliamentary discourse as well as its wider societal impact, in particular with studies that represent diverse parts of society (women, minorities, marginalized groups) and cross-cultural studies (Hughes et al. 2013).

Parliamentary Corpora

The most distinguishing characteristic of records of parliamentary debates is that they are essentially transcriptions of spoken language produced in controlled and regulated circumstances. For this reason, they are rich in invaluable (sociodemographic)

meta-data. They are also easily available under various Freedom of Information Acts set in place to enable informed participation by the public and to improve effective functioning of democratic systems, making the datasets even more valuable for researchers with heterogeneous backgrounds.

This has motivated a number of national as well as international initiatives (for an overview, see Fišer and Lenardič 2018) to compile, process and analyse parliamentary corpora. They are available for most countries within the CLARIN ERIC research infrastructure for language resources and technology, with the UK's Hansard Corpus being the largest (1.6 billion tokens) and spanning the longest time period (1803–2005) while corpora from other countries are significantly smaller (most comprise between 10 and 100 million tokens) and cover significantly shorter periods (mostly from the 1970s onwards).

The Slovene parliamentary corpus SlovParl 2.0 (Pančur 2016) contains minutes of the Assembly of the Republic of Slovenia for the legislative period 1990–1992 when Slovenia became an independent country. The corpus comprises over 200 sessions, almost 60,000 speeches and 11 million words. It contains extensive meta-data about the speakers, a typology of sessions and structural and editorial annotations and is uniformly encoded to the Text Encoding Initiative (TEI) Guidelines, a de-facto standard for encoding and annotating textual data in Digital Humanities. It is available under the CC-BY licence in the CLARIN.SI repository of language resources and via the CLARIN.SI concordancers (Pančur et al. 2017). SlovParl is thus an exemplary corpus but contains material from a quite limited, and not very recent time period. This makes the corpus of limited use for the rich body of research on recent parliamentary activities.

Contemporary Slovenian parliamentary debates are monitored by the analytical tool Parlameter[11] which makes use of linguistic as well as non-linguistic data, such as MPs' attendance and voting results. While this is a very useful tool for journalists and citizen scientists and gives valuable insight into contemporary parliamentary data, its functionality is confined to that of the tool and as such cannot be freely manipulated by scholars according to their specific research needs.

The goal of the research presented in this paper was to convert the Parlameter database into a freely and openly available linguistically annotated corpus enriched with session and speaker metadata, and to showcase the analyses that can be performed on such corpora via open-source tools for corpus analysis. Section 3 gives the basic information on the corpus structure and size, Section 4 presents the analysis of the corpus according to the text and speaker metadata by utilizing some of the best-known corpus analysis techniques, and Section 5 gives some conclusions and directions for further research.

While the focus of the paper is the parliamentary language material which we process with natural language processing and analyse with standard methods from corpus linguistics, the aim of the analysis is to inform media and political studies by transferring the presented methodology into these areas.

1 *Parlameter*, https://parlameter.si.

Corpus Compilation

The data dump from the Parlameter tool consisted of the minutes of the National Assembly of the Republic of Slovenia from its VIIth mandate spanning sessions that started from 2014-08-01 to 2018-05-24 (the complete mandated lasted till 2018-06-22). It was received from the Parlameter API (application programming interface) as a series of JSON files, which were first reorganised into a file containing speaker metadata and a file with the transcriptions of the minutes with speaker identifiers. The speaker metadata contains information about the speaker name and surname, and (for some speakers) their sex, date of birth, education, and party affiliation. The complete speaker metadata is available for the members of the parliament and of the government, but not for, e.g., visiting field experts, representatives of governmental agencies, non-governmental organizations or civil initiatives. This is why the analyses in Section 4 are performed based on the instances for which the metadata is available in the corpus.

The transcriptions contain the ID of the session, name of the session (e.g. "*4. izredna seja*" - *4th extraordinary session*), the date when the session started, and its speeches, each one with the ID of the speaker and a number of segments, roughly corresponding to paragraphs. As discussed below, the transcriptions also contain comments by the transcribers.

Normalisation of Speaker Data

The speaker data was normalised by removing extraneous spaces and removing honorifics (sometimes the name was preceded by, e.g., "*Gospod*" – *Mr.*). Furthermore, in Slovene it is relatively easy to infer the sex from the given name, so we also added sex information to the speakers missing it.

Normalisation of Transcriptions

The JSON dump also contained empty speeches, as well as a significant amount of duplicated speeches. These were removed, as well as extraneous spaces in the text of the transcriptions.

Second, apart from the speeches, the minutes also contained 65,965 comments on verbal and non-verbal behaviour of the speaker or the members of parliament, and there are two types of such remarks. The first are written between slashes and are mostly comments on audible incidents, e.g., */nerazumljivo/ (incomprehensible), /oglašanje iz dvorane/ (comments from the hall), /znak za konec razprave/ (sign for the end of the discussion)*. The second type of comments are written between brackets and mainly denote voting results, e.g., *(nihče), /nobody/, (10 članov) /10 members/, (proti 44) /44 against/*. Both types of comments have been removed from the transcriptions

for the current version of the corpus, as they are not part of the transcription proper and would significantly complicate further processing. Furthermore, the content of the comments is not uniform, with the same information written in various ways (e.g. */smeh/ – laughter, /smeh iz dvorane/ – laughter from the hall, /smeh v dvorani/ – laughter in the hall*), meaning that the values would have to be unified before being converted to appropriate corpus elements.

Linguistic Annotation

In the second stage, the text of the transcriptions was automatically annotated with linguistic information. In particular, the text was tokenised, i.e. split into words, punctuation marks and spaces, and segmented into sentences, which was performed by the ReLDI tokeniser (Ljubešić et al. 2016). Second, the words were part-of-speech tagged and lemmatised, i.e. each word was assigned its context-dependent morphosyntactic description and non-marked form, e.g., the words in "*V naši sredini*" – *In our midst* are assigned the MSDs "*Sl Ps1fslp Ncfsl*" meaning preposition in the locative case; the possessive pronoun in the first person feminine singular locative with a plural owner number; and the feminine common noun in the singular locative, while the lemmas are "*v naš sredina*". The tagging and lemmatisation was performed with the ReLDI tagger (Ljubešić and Erjavec 2016) using its model for Slovene. Finally, the transcriptions were also tagged for named entities, i.e., names identified in the corpus were marked and categorised into five classes, those for persons, locations, organisations, for adjectives derived from a person's name (e.g. "*Cerarjev*" – *Cerar's*), and a miscellaneous category. The named entity annotation was performed with Janes-NER (Fišer et al. 2018).

Corpus Encoding

The corpus is encoded in XML, according to the Text Encoding Initiative Guidelines (TEI Consortium 2017). The complete corpus is stored as one TEI document, which contains its TEI header with the metadata for the corpus, and its text body, containing the transcriptions, one division for each starting date of the sessions; each division is stored as a separate file, giving one root file for the corpus and 525 files for the divisions.

The TEI header contains extensive metadata for the corpus as a whole, e.g., its authors and funders, the source description, the list and numbers of elements used in the corpus, as well as the list of speakers and their metadata. Most metadata is given both in Slovene and English.

As illustrated in Figure 1, the TEI text body date divisions contain a division for each session, and then the utterances for each speaker, each one containing one or more segments, which then contain the annotated transcription.

Figure 1: The TEI encoding of the corpus.

```
<div xmlns="http://www.tei-c.org/ns/1.0" type="date">
 <docDate when="2014-08-26">26.08.2014–</docDate>
 <head>Mandat VII, 26.08.2014–</head>
 <div type="session">
  <head>2. redna seja</head>
  <docDate when="2014-08-26">26.08.2014–</docDate>
  <u xml:id="u529092" who="#spk11">
   <seg xml:id="u529092.seg1">
    <s xml:id="u529092.seg1.1">
     <w lemma="lepo" ana="mte:Rgp">Lepo</w><c> </c>
     <w lemma="pozdravljen" ana="mte:Appmpn">pozdravljeni</w>
     <pc ana="mte:Z">.</pc><c> </c>
    </s>
    <s xml:id="u529092.seg1.2">
     <w lemma="pričenjati" ana="mte:Vmpr1p">Pričenjamo</w><c> </c>
     <w lemma="2." ana="mte:Mdo">2.</w><c> </c>
     <w lemma="seja" ana="mte:Ncfsa">sejo</w><c> </c>
     <w lemma="kolegij" ana="mte:Ncmsg">Kolegija</w><c> </c>
     <w lemma="predsednik" ana="mte:Ncmsg">predsednika</w><c> </c>
     <name type="org">
      <w lemma="državen" ana="mte:Agpmsg">Državnega</w><c> </c>
      <w lemma="zbor" ana="mte:Ncmsg">zbora</w>
     </name>
     <pc ana="mte:Z">.</pc>
    </s>
```

Corpus Size

Some basic statistics regarding the corpus are given in Table 1. In total, the Parlameter corpus contains 371 sessions (as distinguished by their title) which spanned over 525 days, i.e., 1.4 days per session on average. If we count distinct sessions that started on a given day, the corpus contains 1,338 such sessions. The VIIth mandate of the parliament heard 1,981 speakers who gave 133,287 speeches which contain almost 35 million words, i.e., 67 speeches per speaker and 260 words per speech on average. Due to a number of factors, such as different roles of the speakers in the parliament, the distribution is, of course, far from uniform, e.g., there is one speaker that gave 14,616 speeches, while 711 speakers gave only one speech.

Table 1: Basic statistic of the Parlameter corpus.

Tokens	40,987,516
Words	34,882,499
Sentences	1,833,147
Utterances	133,287
Speakers	1,981
Sessions on date	1,338
Dates	525
Sessions	371

Availability of the Corpus

The Parlameter corpus is available through CLARIN.SI. CLARIN is the European research infrastructure for language resources and technologies, which makes digital language resources available to scholars, researchers, students and citizen-scientists from all disciplines, especially in the humanities and social sciences, through single sign-on access. CLARIN offers long-term solutions and technology services for deploying, connecting, analysing and sustaining digital language data and tools. CLARIN is organised as a network of national centres, with CLARIN.SI covering Slovenia. CLARIN.SI[2] offers, inter alia, two concordancers for on-line corpus exploration, and a repository of language resources and tools, intended for their long-term archiving together with support for different types of licences and an unambiguous way for others to cite these resources, using Handle persistent identifiers. The landing page of each resource also gives a cross-reference to the concordancers for the particular corpus, and vice-versa. The repository also exposes its metadata, which is being harvested by a number of other services.

The Parlameter corpus is available through both CLARIN.SI concordancers, as well as for download from its repository, both as a TEI document and in the simpler vertical file format, under the liberal Creative Commons – Attribution-ShareAlike (CC BY-SA 4.0) licence (Dobranić et al. 2019). In this way we hope to raise interest among other researchers to explore the corpus and make use of it in their research.

2 *CLARIN Slovenia*, http://www.clarin.si/info/about/.

Corpus Analysis

By using the CLARIN.SI NoSketch Engine concordancer,[3] we demonstrate the potential of the basic corpus analysis techniques (Gorjanc and Fišer 2013) for politology, history and other related humanities and social sciences disciplines that base their research on large volumes of language data. *Concordances* are lists of all examples of the search word or phrase from a corpus which are shown in the context they were used in and are equipped with the available metadata. *Wordlists* are comprehensive summarizations of the language inventory in the corpus, organized by frequency or alphabetically. *Collocations* are partly or fully fixed multi-word expressions which have become established through usage. *Keywords* are words which appear in the focus corpus more frequently than they would in the general language. Combined with the available text and speaker metadata, such as date, speaker gender or political affiliation, they provide a powerful analytical tool for discovering the commonalities and specificities of the linguistic footprint and trends by different types of speakers in the parliament as will be shown in the rest of this section.

Production Volume and Vocabulary Size

As already presented in Table 1, the corpus contains nearly 41 million tokens or 35 million words. noSketch Engine also offers the lexicon size of the corpus, as given in Table 2, which shows that the corpus contains approximately 263,000 different word forms (so, inflected words, e.g., *Slovenije*) and over 104,000 different lemmas (so, base forms of words, e.g., *Slovenija*), and 1,080 different morphosyntactic tags (e.g.,*Verb main present second plural*). However, it should be noted that both lemmas and the tags are automatically assigned, so they also contain some annotation errors: the accuracy of morphosyntactic tags is around 94%, the accuracy of lemmas is above 99%.

Table 2: Lexicon sizes of the Parlameter corpus.

Unique words	263,007
Unique lemmas	104,247
Unique tags	1,080

While the corpus contains parliamentary debates from the period 2014-2018, 62% of the material was recorded in 2015 and 2016. Given the parliamentary term, which lasted from 1 August 2014 to 14 April 2018, it is interesting to observe an 8% smaller production in 2017 compared to the year before since the last year of the term would be expectedly the busiest in order to wrap up the workplan and set the ground for a new election cycle.

3 *NoSketch Engine @ CLARIN.SI*, https://www.clarin.si/noske/.

Table 3: Distribution of text quantity by year in Parlameter.

Year	No. of tokens	% of tokens	Rel. freq.
2014	3,759,110	9%	91,714
2015	12,441,754	30%	303,550
2016	13,270,257	32%	323,763
2017	9,944,401	24%	242,620
2018	1,571,994	4%	38,353
Total	40,987,516	100%	1,000,000

Morphosyntactic Specificities of the Language in ParlaMeter

We performed a basic analysis of the morphosyntactic annotations of the corpus in form of the most significant differences in their frequencies between the Gigafida reference corpus of Slovene[4] and the Parlameter corpus, which are given in Table 4.[5]

Table 4: Most salient differences in morphosyntactic descriptions between Gigafida 2.0 and Parlameter.

Gigafida	Parlameter
Residual web	Pronoun personal first singular nominative
Numeral roman cardinal	Verb main present second plural
Adjective possessive positive masculine singular instrumental	Pronoun personal second masculine plural nominative
Auxiliary infinitive	Pronoun possessive first feminine singular genitive singular
Adjective possessive positive masculine plural genitive	Verb main present first plural -Negative
Adjective possessive positive masculine singular locative	Verb main present second plural -Negative
Adjective possessive positive neuter singular locative	Pronoun demonstrative neuter plural accusative
Pronoun possessive third masculine singular accusative dual	Pronoun personal first singular accusative
Adjective possessive positive masculine singular nominative -Definiteness	Verb main present first singular

4 For this comparison we used the deduplicated version of Gigafida 2.0. At the time of writing, this corpus was newly made and does not yet have a reference publication. It is, however, freely available for searching and analysis at https://www.clarin.si/noske/.

5 The morphosyntactic tags are given here in their expanded form to aid understanding. The reference to these morphosyntactic descriptions is given in http://nl.ijs.si/ME/V6/msd/html/msd-sl.html.

Gigafida	Parlameter
Pronoun possessive third feminine plural locative singular masculine	Verb main present first singular
Adjective possessive positive masculine plural nominative	Pronoun demonstrative masculine singular dative
Noun proper feminine plural dative	Pronoun indefinite feminine singular genitive
Numeral letter ordinal neuter plural genitive	Pronoun indefinite masculine singular accusative
Pronoun personal first dual accusative	Verb auxiliary present second plural -Negative
Pronoun personal first dual dative	Verb auxiliary future first singular -Negative
Noun proper neuter singular instrumental	Pronoun personal first masculine plural nominative
Adjective possessive positive feminine singular locative	Verb auxiliary present second plural +Negative
Pronoun personal second singular accusative bound	Verb main present first plural
Pronoun personal third masculine dual dative +Clitic	Pronoun indefinite feminine singular accusative
Adjective possessive positive masculine plural locative	Pronoun demonstrative feminine plural accusative

The results show that the parliamentary speeches, as expected, contain more present tense verb forms, especially in the first and second person singular or plural (e.g., *imamo – we have, pozdravljam – I greet, zaupate- you trust*), as well as personal and demonstrative pronouns, the former most prominently as the first person singular personal pronoun *(jaz – I)*.

On the other hand, the parliamentary proceedings do not contain URLs or Roman numerals. More interestingly, they also contain significantly fewer possessive adjectives (e.g. *torkovim – Tuesday's*) and pronouns (*njun – theirs*$_{[dual]}$), proper names, numerals, personal pronouns in the dual number (*naju – us two*), or in second person singular accusative (*nate – to you*) than general Slovene.

Language and Gender in Parlameter

Gender is recorded for all but one speaker in the corpus.[6] In total, 1,965 speakers are represented, 62% of which are male and 38% female. Interestingly, the contribution from the speakers is not proportionate to the distribution according to their gender,

6 This missing information is due to errors in input metadata records, which will be improved in the next version of the corpus.

with the male speakers contributing 71% of the tokens in the corpus and the female speakers 29%. On the speech level the difference is even more pronounced as the male speakers delivered 73% of the speeches while female speakers only 27%, indicating that, on average, the speeches given by female speakers were somewhat longer than those by male speakers.

Table 5: Distribution of speakers and text production by gender in Parlameter.

Gender	No. of speakers	% of speakers	No. of tokens	% of tokens
Female	747	38%	29,147,871	71%
Male	1217	62%	11,838,913	29%
Unknown	1	0%	732	0%
Total	1965	100%	40,987,516	100%

Table 6, which lists top-ranking 10 female and male speakers and their production in terms of tokens, shows that the most prolific male speakers produced nearly twice as much material as their female counterparts. Overall, all top 10 speakers except one (Miha Kordiš, male, the Levica party) have a leading role in one or more parliamentary or governmental bodies, including 2 ministers, both of which are female, 2 opposition deputy group chairs, who are both male, and the Chair of the National Assembly who is also male. Based on their roles in the parliament or the government, top-ranking speakers represent issues on culture, corruption, judiciary, finances, agriculture, foreign policy, education and infrastructure. In terms of political orientation, the largest opposition party SDS is best represented with 5 top-ranking male and 3 female speakers, including chair and vice-chair of their deputy group. Among the top-ranking female speakers, the entire political spectrum is represented while male speakers from the SD and DeSUS parties do not make the list, and the SMC party is only represented by the Chair of the National Assembly whose role is most likely predominantly procedural, not to promote the party agenda.

Table 6: Top-ranking 10 female and male speakers and their text production in Parlameter.

Female	Party affiliation // Role	Tok. %	Male	Party affiliation // Role	Tok. %
Anja B. Žibert	SDS // Chair of the Culture Committee	698,883 6%	Jožef Horvat	NSI // Chair of the Foreign Policy Committee; Chair of the Deputy Group NSI	1,141,778 4%
Jelka Godec	SDS // Chair of the Inquiry Commission on the Misuse Practices in Healthcare	530,029 4%	Jani Mödern-dorfer	ZAAB // Chair of the Inquiry Commission on bank money laundering; Vice-chair of the Election Committee	1,062,546 4%

Female	Party affiliation // Role	Tok. %	Male	Party affiliation // Role	Tok. %
Iva Dimic	NSI // Vice-chair of the Judiciary Committee	509,101 4%	Franc Trček	Levica // Vice-chair of the Infrastructure Committee; Vice-chair of the Inquiry Commission on bank money laundering	1,060,399 4%
Alenka Bratušek	ZAAB // Vice-chair of the Public Finances Committee; Vice-chair of the Deupty Group ZAAB	483,171 4%	Milan Brglez	SMC // Chair of the National Assembly; Chair of the Constitution Committee	948,334 3%
Violeta Tomić	Levica // Vice-chair of the Agriculture Committee	446,460 4%	Vinko Gorenak	SDS // Vice-chair of the Deputy Group SDS	788,678 3%
Eva Irgl	SDS // Chair of the petition committee	439,042 4%	Franc Breznik	SDS // Vice-chair of the Election Committee	763,437 3%
Urška Ban	SMC // Chair of the Finances and Monetary Policy Committee	382,425 3%	Jože Tanko	SDS // Chair of the Deputy Group SDS	752,130 3%
Mateja V. Erman	Minister of Finance	381,604 3%	Andrej Šircelj	SDS // Chair of the Public Finances Committee	721,135 2%
Bojana Muršič	SD // Vice-chair of the National Assembly, Vice-chair of the Education Committee	366,547 3%	Tomaž Lisec	SDS // Chair of the Agriculture Committee	707,666 2%
Julijana B. Mlakar	DeSUS // Minister of Culture; Vice-chair of the Foreign Policy Committee	308,355 3%	Miha Kordiš	Levica	676,717 2%

In order to compare the topics discussed by female and male speakers in the Slovene parliament, we analysed their 100 top-ranking key lemmas, where we used the corpus of all female speakers as the target corpus against the reference corpus of all male speakers in the Parlameter corpus, and vice versa, so the two lists display the distinguishing features of each of the groups. By observing their contexts via concordances, we manually classified them into one of the 13 topics represented by the ministries in the Slovenian government:

- *agriculture, forestry and food*
- *culture*
- *defence*
- *economy and technology*
- *education, science and sport*
- *environment and spatial planning*
- *finance*
- *health*
- *foreign affairs*
- *infrastructure*
- *interior*
- *justice*
- *labour, family and social affairs*
- *public administration*

In addition, we introduced 4 additional categories for words that could not be classified into any of the topics above:

- *interaction/procedural* for keywords which referred to other people attending the session (e.g., references to names of other speakers, *predsednik – chairman*) or expressed procedural matters during the session (e.g., *prisotni – present, dobrodošli – welcome*)
- *style* for keywords which were either distinctly colloquial or distinctly formal and were frequently used only by a single or very few speakers in order to achieve a special effect (e.g., *penez*, a very informal expression for money, *šiht*, a very informal expression for job)
- *ideology* for keywords which were used to ideologically label an individual speaker or a group of speakers (e.g., *levičarski – leftist, kapitalizem – capitalism*)
- *multiple* for keywords which were used in several topics (e.g., *zgodnji – early, fantastičen – fantastic*).

As can be seen from Table 7, the most frequent topics among the female speakers are *health* (35) and *labour, family and social affairs* (33), which are followed by *public administration* (13) and *education, science and sport* (8). Most of the 100 top-ranking keywords uttered by male speakers, on the other hand, could not be classified into a single topic because they were used either to achieve a *stylistic effect* (24), were general words that were used in *multiple topics*, such as descriptive adjectives or legal terms (22), or *ideological expressions* (6), all of which indicate a more discursive, debating style of the male speakers, but could also stem from the fact that the leading roles in that term were predominantly held by male members of parliament.[7] Despite being much more infrequent than in the female part of the corpus overall, the most

7 This problem could be avoided by removing outliers regarding production in the dataset before performing the analyses. But our goal here was to present the complete corpus and demonstrate the basic corpus analysis techniques.

frequently represented specific topics by male speakers are *infrastructure* (9), *interior* (6), *agriculture, forestry and food* (5), and *defence* (5), suggesting a significant difference in the roles and interests of male and female speakers in the Slovene parliament.

Table 7: Topics of 100 top-ranking keywords of female and male speakers in Parlameter.

Topics – female	Freq.	Topics – male	Freq.
health	35	style	24
labour, family & social affairs	33	multiple	22
public administration	13	infrastructure	9
education, science & sport	8	interior	6
interaction/procedural	3	ideology	6
multiple	3	interaction/procedural	5
environment & spatial planning	1	agriculture, forestry & food	5
agriculture, forestry & food	1	defense	5
culture	1	foreign affairs	4
finance	1	finance	4
economy & technology	1	justice	3
Total	100	Total	100

Illustrative examples of the 10 top-ranking female- and male-specific keywords with a manually assigned topic are listed in Tables 8 and 9.

Table 8: Most frequent keywords, topics and word type among female speakers in Parlameter. N stands for nouns, Adj for adjectives, and NP for proper nouns (names).

Lemma – English translation	Topic	PoS	Freq.	Freq_ref	Score
rejništvo – fostercare	labour, family & social affairs	N	264	59	7.7
mark – mark	health	PN	155	29	7.1
enostarševski – single-parent	labour, family & social affairs	Adj	167	38	6.6
roditeljski – parent	labour, family & social affairs	Adj	169	39	6.5
medical – medical	health	PN	128	26	6.2
plazma – plasma	health	N	82	9	6.1
pacientov – patient's	health	Adj	282	97	5.7
zaznamba – notice	public administration	N	155	43	5.7
žilen – stent	health	Adj	518	213	5.4
duševen – mental	health	Adj	393	156	5.4
nasilnež – violent person	labour, family & social affairs	N	98	21	5.4

Table 9: Most frequent keywords, topics and word type among male speakers in Parlameter.

lemma – English translation	category	PoS	Freq_ref	Score
penez – inf. money	finance	N	0	13.2
navsezadnje – nevertheless	multiple	Adv	90	8.4
kubik – cubic	agriculture, forestry & food	N	10	7.8
islam – Islam	interior	N	6	6.4
levičarski – leftist	ideology	Adj	2	6.2
navzoč – present	interaction/procedural	Adj	211	6.0
avtošola – driving school	infrastructure	N	1	5.8
socialist – socialist	ideology	N	25	5.5
svojevrsten – peculiar	multiple	Adj	16	5.4
e-klopa – e-bench	interaction/procedural	N	1	5.3
prečenje – crossing	style	N	3	5.2

That the nature and style of male speeches is quite different from the female ones can also be seen from the analysis of the morphosyntactic types of 100 highest-ranking keywords for male and female speakers. While nouns are the most frequent category and are used equally frequently by both male and female speakers (44%), many more adjectives were found among the female top-ranking keywords (33% vs. 16%), while the male keywords had more adverbs (11% vs. 4%) and verbs (9% vs. 2%), which again could be related to the roles of the speakers in the parliament. However, given the results of our preliminary work on this dataset (Ljubešić et al. 2018), during which we removed the speakers that produced most of the linguistic material from the analysis, we see similar trends both in the gender-dependent keyword and morphosyntactic analysis, and are therefore rather in favour of accepting the observed differences as impact of gender and not role.

Language and Party Affiliation in Parlameter

Affiliation is recorded for only 113 speakers out of the 1982, however, these are responsible for 79% of the tokens in the corpus. Affiliation is considered as either deputy group membership or a role in the government, where it must be noted that in this version of the corpus the metadata reflect the situation at the beginning of the term and does not keep track of party membership transfers or resignations of ministers or members of parliament. Also, when elected members of parliament were later appointed as ministers, the metadata record only their party affiliation and records as ministers only those who were appointed without being first elected to the parliament. To facilitate more fine-grained and accurate use of the corpus in political science or contemporary history, we plan to refine the metadata for the next release of the corpus, adding also the MP's membership in the working bodies of the National Assembly,

etc. Also, the metadata in the current version of the corpus do not flag the independent members of parliament who do not belong to any of the parliamentary parties and operate in the Independents deputy group, which is why they are not included in our analysis.

As Table 10 shows, the most prolific deputy group is the largest opposition party Slovenian Democratic Party (SDS), whose 20 members contributed nearly 10 million tokens or 30% of the corpus. SDS is followed by the main governing party, Party of Modern Centre (SMC), whose 42 members contributed 7 million tokens or 22% of the corpus. It is interesting to note that in terms of the volume contributed to the corpus on one side and the number of speakers on the other, that this party was the least productive among the main parties, with a ratio of the percentage of tokens to the percentage of speakers (i.e., the relative token to speaker ratio) of 0.54, which means that this party generated a little bit more than a half of the material that would have been expected given their number of speakers and the overall activity of all the speakers. The Left (Levica) and New Slovenia (NSi) rank third and fourth, despite the fact that they had only 6 members each in the parliament, making them the most productive parties with a relative token to speaker ratio of 1.83 and 1.66. The Democratic Party of Pensioners of Slovenia had as many as 12 elected MPs but contributed 1 million tokens less than the two previous parties, which makes them the second least productive party with a relative token to speaker ratio of 0.67.

Table 10: Distribution of speakers and text production by party affiliation in ParlaMeter with speakers with unknown affiliation removed.[8]

Affiliation	No. of speakers	% of speakers	No. of tokens	% of tokens
Slovenian Democratic Party Deputy Group (SDS)	20	20%	9.516.651	30%
Party of Modern Centre Deputy Group (SMC)	42	41%	7.162.719	22%
The Left Deputy Group (Levica)	6	6%	3.438.194	11%
New Slovenia – Christian Democrats Deputy Group (NSI)	6	6%	3.370.131	10%
Social Democrats Deputy Group (SD)	9	9%	2.533.019	8%
Democratic Party of Pensioners of Slovenia Deputy Group (DeSUS)	12	12%	2.435.884	8%
Party of Alenka Bratušek Deputy Group (SAB)	4	4%	1.876.294	6%
Italian and Hugarian National Minorities Deputy Group (IMNS)	2	2%	117.709	0%
Government	1	1%	1.765.374	5%
Total	102	100%	32.215.975	100%

8 The number of speakers per party is calculated from the ParlaMeter dump and deviates slightly from the official member number due to different handling of speakers with multiple roles.

Next, we performed a manual analysis of the 100 top-ranking keywords of each political party against the rest of the corpus. These analyses display the distinct properties of one party that are not shared by other parties. Using the concordances, we classified the keywords into the same categories as in Section 4.1, the results of which are summarized in Tables 11 and 12.

Table 11: Topics of 100 top-ranking keywords of party members in Parlameter.

Topics	SMC	DeSUS	SD	SDS	NSi	Levica	SAB
agriculture, forestry & food	0	0	**34**	0	**27**	0	0
culture	0	3	0	0	0	1	0
defense	0	0	**21**	5	0	0	1
economy & technology	0	0	5	1	**11**	**13**	1
education, science & sport	0	0	0	0	0	0	4
environment & spatial planning	0	0	3	0	6	1	0
finance	0	2	2	0	6	1	1
foreign affairs	0	5	0	2	4	3	0
health	0	3	0	8	1	0	5
ideology	0	0	0	**15**	3	9	0
infrastructure	1	0	2	0	7	1	1
interaction/procedural	**99**	**61**	14	**17**	10	4	14
interior	0	0	0	3	0	3	5
justice	0	1	1	8	0	0	0
labour, family & social affairs	0	**13**	3	1	4	13	3
multiple	0	2	6	13	8	17	29
public administration	0	2	0	5	2	1	7
style	0	8	9	**22**	11	**33**	**29**
Total	100	100	100	100	100	100	100

Unsurprisingly, due to the role of the main governing party SMC, practically all their top-ranking keywords are interactional elements with the other speakers or have a procedural nature (e.g., *navzoč – present, glasovanje – voting, amandma – amendment*). That DeSUS is a single-issue party can be seen from their keywords, which, apart from a surprisingly high proportion of interactive keywords, belong almost exclusively to the semantic field of retirement and pension (e.g., *regres – holiday pay, valorizirati – to revalue, gmoten – material*). Interestingly, even the topics of foreign affairs and culture are nearly completely absent from their keyword list, despite the fact that these ministers came from their party, suggesting that these topics are more or less evenly shared with other parties. SD, the third coalition party, clearly display their priority areas of agriculture, forestry and food (e.g., *teran – Teran wine, fermentiran – fermented, kmetovati – to farm*) and defence (e.g., *vojakinja – female soldier, neeksplodiran – unexploded, strelivo – ammunition*), which can be traced back to their ministers.

The largest opposition party SDS stands out from the rest by the amount of ideological keywords identified among the top-ranking keywords (e.g. *tranzicijski – transitional, totalitarizem – totalitarism, lustracija – lustration*). NSi and Levica, the opposition parties with the same number of MPs but from the opposite ends of the political spectrum, both address the widest variety of issues (their keywords were classified into 13 out of 18 topics). The topics with nearly equal number of completely opposite keywords are economy and technology (e.g. *soupravljanje – co-management* for Levica vs. *espejevec – private entrepreneur* for NSi). While NSi mostly talks about the local issues related to their constituencies (e.g. *samooskrba – self-sufficiency, posekan – cut down, obdelovati – farm*), Levica stands out by signature stylistic devices which range from very informal (e.g. *šlamastika – pickle, gazda – informal for master, nabijati – to bang on*) to highly elevated registers (e.g. *nemara – perhaps, onkraj – beyond, ducat – dozen*) and displays the largest proportion of ideological vocabulary next to SDS (e.g. *tovarišica – camerade, revizionizem – revisionism, imperializem – imperialism*). SAB seems to stand out by a predominantly (local) administrative/procedural/governance vocabulary (e.g. *proporcionalen – proportional, odpoklic – recall, dvokrožen – double-ballot*) as well as a discursive, informal style of distinctly negative sentiment, which is characteristic of one of their members Vinko Möderndorfer (e.g. *rešpektiram – honour, kozlarija – nonsense, zmazek – disaster*).

Table 12: 100 top-ranking keywords per political party, taking into account lowercased lemmas, computed against the rest of the Parlameter corpus and sorted according to their keyness score.

SMC	navzoč, e-klopa, udis, roberto, prekinjen, podprogram, prehajati, lipicer, kustec, katerim, grebenšek, h, battelli, epi, stanujoč, obveščati, krajnc, zaključevati, predajati, pričenjati, sodin, porotnica, simona, franc, glasovati, obrazložitev, moderen, kolegij, tanko, postopkovno, potisek, končevati, nuklearen, brezpredmeten, ep, jernej, dneven, počkaj, glasovnica, mandatno-volilen, vojko, jožef, trček, bojan, neusklajen, tilen, prelog, ustavnorevizijski, odločanje, arko, nadomeščati, he, branislav, matej, jože, glasovanje, prvopodpisan, e-klop, glas, dopolnjen, porotnik, terminski, vložen, simono, franca, pogačnik, erman, ugotavljati, klanjšček, smc, stebernak, nepovezan, jana, žibert, bien, matjaž, šircelj, fajt, postopkoven, lilijana, skrajšan, monetaren, prekinjati, poslovniški, matičen, bah, mag., marinka, šergan, lenča, vraničar, izvolitev, karlovšek, razpravljavec, predstavnica, razširitev, anita, amandma, nadomeščanje, zame
DeSUS	meglič, črnak, pripadajoč, desus, pogačar, dasiravno, vukov, valenca, požun, inferioren, upajoč, möderndorfer, pregrešiti, divjak, valorizacija, korva, rezime, kkr, kuzmanič, marijan, upokojen, vuk, mehčati, pojbič, košnik, bližnjevzhoden, zaposlovalen, punkcija, žmavc, milojka, zaporedno, celarc, konzularen, xv., marija, kolar, bačič, erika, grošelj, rubelj, minski, lukić, rudarski, zadržanost, mirjam, godec, valorizirati, sng, tašner, kušar, brinovšek, invalid, zamrznitev, tedaj, dvoživkarstvo, nina, pirnat, dekleva, merše, federacija, nada, klanjšček, protiukrep, jelka, ogrizek, gmoten, kisikov, ivo, majcen, izvoliti, iva, dimic., modifikacija, ljubič, žan, upokojenec, prikrajšanje, prečitati, šimenko, jasna, izplačevanje, zipro, korpič, antonija, premožen, sapa, voljč, suzana, dimic, vesni, lukič, zdravko, irena, teja, sluga, regres, ruše, janja, razparava, trivialen
SD	izčistiti, genetsko, izčiščen, vezava, surov, demokrat, vojakinja, gorsko-hribovski, travinje, potočan, vadišče, razprodati, hip, služenje, hišniški, faktorski, pripadnica, stiskanje, zmogljivost, omd-, kočevski, anhovo, vrtojba, peterica, mineralen, maji, krušen, kmetica, ciolos, vklop, deti, socialdemokratski, formacijski, teran, selnica, kloniran, urszr, obramben, salonit, radeče, mlekarna, neperspektiven, marjana, popolnjevanje, omd, odzivanje, vrtnina, vselej, zorganizirati, vikariat, eutm, pokolp, govedo, rogaška, klirinški, razprodaja, surovina, ksenija, vinko, izčiščevati, konzumen, refundirati, pripadnik, neeksplodiran, social, uokviriti, žito, kfor, prebroditi, konvergenca, grajski, brecelj, hogan, administriranje, trader, kočevsko, h4, primož, korenjak, bržkone, kmetovati, obrtništvo, vojska, strelivo, poveljevanje, snežnik, plasiran, gorsko, refundacija, hribovski, proizvodnja, subvencijski, dacian, missing, kmetija, opazovati, voditeljstvo, kramar, fermentiran, viher
SDS	islam, fišer, mark, svinjarija, levičarski, odnosno, medical, kb, demokratski, odnosen, lenart, zemljarič, kučan, zalar, bordojski, kb1909, morišče, zločin, iznenada, velikanski, tomos, kangler, patria, multikulti, masleša, prvorazreden, škrlec, udba, stožice, tranzicijski, šef, praprotnik, moralno-etičen, ilegalno, zločinski, bomben, peticija, porsche, srebrenica, cener, umor, totalitaren, pokrasti, totalno, genocid, drugorazreden, tamle, erdogan, judikat, vega, ribičič, privilegiranec, komunističen, razorožitev, varnostnoobveščevalen, žilen, opornica, indičen, škandal, ornik, lustracija, poljanski, posavje, počenjati, furlan, pobiti, sevnica, ubog, janković, krkovič, npu, deček, opran, bojda, blamaža, lopov, toplak, kerševan, slikati, bmw, veselo, amen, totalen, komunizem, totalitarizem, obsoditi, preiskati, bedarija, udbovski, pomorjen, turnšek, vladavina, zlagati, šoping, vpiti, ukc, avion, klemenčič, koruptiven, neumnost

NSI	komunalno, socialno-tržen, marn, božičnica, zidanica, egalitaren, krščanski, espejevec, fantastičen, ekstrapolacija, planšarija, medparlamentaren, kamnik, demografija, kapica, bundestag, podonavski, bajuk, samoprispevek, vinogradnik, razlastiti, vipavski, prijateljstvo, kanalizacija, aksiom, pomurje, bogataš, ferenc, parcelacija, optimirati, oljčnik, komenda, polnost, vrtalec, ozp, pomurski, ikt, simulirati, dimniški, parlamentarec, podčrtovati, artikulirati, obžalovati, omizje, cerknica, polčas, ginijev, zbirno-reciklažen, brutalno, prekladanje, širokogruden, absorpcijski, šinko, dolenjsko, lestev, vodovod, rodnost, traktor, notranjska, opn, posekan, vinograd, zaraščati, odvajanje, loža, kristjan, davno, regresen, lovrenčič, firefox, parcela, akrapovič, obdelovati, obratovalnica, zpn, terezija, mihael, odlašati, peskovci, vamp, notranjski, ovs, copatek, veselica, upniški, penzija, hala, digitalen, goljuf, identifikacijski, mohar, postoriti, goveji, prirasti, splačati, samooskrba, prazniti, odstaven, todorić, pozor
Levica	penez, tuliti, vračljivost, ubesedovati, onkraj, bajta, neoliberalen, prečiti, nemara, ducat, socialist, delavski, imperialističen, zvrniti, desnica, navsezadnje, blazen, sociolog, šiht, soupravljanje, zategovanje, mandarin, kapitalizem, strokovec, šlamastika, blazno, kapitalističen, tovarišica, ubesedovanje, revizionizem, prekarnost, vzdržan, gazda, profit, sodržavljanka, izkoriščevalski, represija, protisocialen, nabijati, prekaren, metafora, soodločanje, periferen, agregaten, cinkarna, rezilen, mezda, amandmiranje, demokratizacija, ips, efektivno, natov, levica, belokranjec, bučka, zaposlovalec, izhajajoč, reven, požegnati, profiten, marof, ics, minimalec, podrejati, imperializem, kapitalist, silno, prekarizacija, odpustek, sodržavljan, noveliranje, versus, zvo, bolgarski, zastraševanje, informatičen, metaforično, režati, razreden, ciničen, striči, ropotati, korporacija, rasizem, redistributiven, pregrevanje, trade, rez, omv, prekeren, deregulacija, štacuna, grosist, znoreti, penzion, oligopolen, jahati, fevdalizacija, sočasno, prečenje
SAB	svojevrsten, večnost, mvk, pooblaščati, that's, diskvalifikacija, prekleto, bla, resnica, fakt, naglas, odpoklic, zavezništvo, minis, četrten, trapast, istrabenz, zasebništvo, zamah, dvokrožen, ramšakov, diskvalificirati, športnica, drk, štos, cetera, ups, nedostojno, redarski, strojan, nijz, proporcionalen, ma, evtanazija, zanič, bloudkov, etc, mv, vsakič, naturalizacija, zamera, nor, listnica, smešiti, dispečiranje, diskusija, strašansko, nefer, diskutirati, regres, sprevržen, r., zavrtanik, večen, hiv, nekorektno, ubežati, imperativen, presedan, prastrah, dinozaver, halo, ekstremističen, rimskokatoliški, mvk-, namenoma, zmazek, gedrih, somalijski, zamahniti, nonstop, kostanjevec, policaj, domišljati, prohibicija, znakoven, paradoks, barantati, et, hecen, močvirnik, avans, nametati, preprosto, prepričevati, podžupan, traparija, kričati, ekstra, non-stop, telovadba, stefanovič, el-zoheiry, ničkolikokrat, kozlarija, prvenstvo, boh, domišljija, rešpektiram

The Zeitgeist of ParlaMeter

Finally, we observe the zeitgeist of the Parlameter corpus by comparing it with its older and smaller cousin, the SlovParl corpus, which contains material from the period of Slovenia's independence (1990–1992). First, we created keyword lists with each of the two corpora acting as a focus and a reference corpus. We then manually classified 100 top-ranking keywords into the same categories as in Section 4.1, with the following additional categories:

– abbreviations (*etc., Mr.*), which were in use in the SlovParl but are no longer the convention in the ParlaMeter transcriptions of the parliamentary sessions
– IT vocabulary (*internet, web*), which at the time of SlovParl was not yet widespread.

If we disregard the differences in the mentions of the active politicians in the two periods, which are the most frequent category, most of the top-ranking keywords in both corpora belong to procedural and legal issues, which are clearly different in a newly established state and a state integrated in the EU (see Tables 13 and 14). Apart from that, many more topics are identified in the Parlameter corpus, such as economy and technology, foreign affairs and health, which again is not surprising as a well-established state will need to take care of a full spectrum of issues.

Table 13: Topics of the 100 top-ranking keywords in Parlameter and SlovParl.

Topic	ParlaMeter	SlovParl
abbreviation	0	3
defence	0	1
economy & technology	6	2
education	1	0
environment & spatial planning	2	0
finance	12	7
foreign affairs	4	0
health	4	0
multiple	0	1
informal vocabulary	2	0
infrastructure	1	0
interior	2	0
it vocabulary	2	0
justice	1	0
labour, family & social affairs	3	0
legal/procedural	14	21
politician/party	46	65
Total	100	100

Table 14: 100 top-ranking keywords in Parlameter contrasted against SlovParl and vice versa.

ParlaMeter	evro, eu, desus, smc, cerar, sdh, dutb, möderndorfer, trček, bratušek, sds, gorenak, spleten, mandatno-volilen, deležnik, koalicijski, kordiš, anja, matej, direktiva, postopkovno, kpk, okoljski, kohezijski, javnofinančen, tonin, bdp, veber, naročanje, korupcija, bah, jani, levica, nlb, unija, tanko, migrantski, povprečnina, vatovec, čakalen, pojbič, migrant, varuhinja, prikl, žnidar, šircelj, varuh, zujf, teš, violeta, tomić, mahnič, ddv, digitalen, han, istospolen, lisec, telekom, vrtovec, dars, žibert, novela, globa, zorčič, vajeništvo, godec, trošarina, čuš, okrožen, internet, prvopodpisan, schengenski, matić, trajnosten, gašperšič, jurša, podneben, dz, lipica, lah, podizvajalec, žan, uredba, blagajna, okej, verbič, ferluga, dobovšek, mramor, računski, vraničar, zakonik, ljudmila, nevladen, postopkoven, preiskovalen, direktorat, hanžek, muršič, irgl
SlovParl	delegat, oz., glavič, družbenopolitičen, gros, dinar, republiški, usklajevalen, din, skupščinski, starman, zakonjšek, alinea, vzdržati, potrč, vzdržan, kolešnik, izvršen, lukač, sklepčnost, pintar, npr., navzočnost, buser, arzenšek, feltrin, atelšek, liberalno-demokratski, smole, razpravljalec, školč, zvezen, schwarzbartl, delegatski, tomšič, zagožen, železarna, jakič, gošnik, skupščina, polajnar, tomažič, muren, štefančič, lastninjenje, deviza, zlobec, šter, demos, dretnik, kreditno-monetaren, sdp, čimprej, nabornik, devizen, marka, delegatka, sekretariat, bekeš, deželak, klavora, peterle, črnej, halb, kreft, šonc, lokar, gradišar, šeligo, juri, perko, sfrj, voljč, požarnik, semolič, volilec, kramarič, bučar, plebiscit, dvornik, tomše, grašič, tolar, starc, pregelj, podobnik, pozsonec, balažic, g., moge, medzborovski, jaša, razdevšek, rojec, šetinc, urbančič, lavtižar-bebler, vivod, anka, šešok

To illustrate differences in the zeitgeist of both corpora, we extracted the strongest collocations of the following 3 expressions, which are frequent in both corpora, taking into account the collocation candidates that appear at least 5 times immediately next (left or right) to the headword, and analysed the first 50 collocation candidates:

– adjective *južen – southern*,
– noun *kriza – crisis*, and
– verb *sprožiti – trigger*.

Table 15: Comparison of collocations of *južen*, *kriza* and *sprožiti* in SlovParl and ParlaMeter. Topics or morphosyntactic categories are indicated in square brackets, and new collocations in Parlameter are highlighted in bold.

	SlovParl	**ParlaMeter**
južen	178 (14.03 per million) - [GEOGRAPHY]: koreja, primorska, amerika - [CONCRETE]: meja, železnica - [METAPHORICAL]: trg, del, stran, republika	910 (22.20 per million) - [GEOGRAPHY]: **afrika**, koreja, **sredozemlje**, amerika, tirolska, sudan, **tirolec**, **koroška, italija, evropa, nemčija, slovenija** - [CONCRETE]: meja, **obvoznica, tok, sadje, odsek,** železnica, **ulica** - [METAPHORICAL]: **sosedstvo, soseda, sosed, soseščina,** del, trg, **projekt, stran, država, republika**
sprožiti	548 (43.19 per million) - [CONCRETE]: spor, postopek, proces, interpelacijo, arbitražo - [METAPHORICAL]: reakcijo, polemiko, akcijo, mehanizem, pobudo, vprašanje, diskusijo, zahtevo, spremembo, razpravo, zadevo	1,569 (38.28 per million) - [CONCRETE]: postopek, spor, **preiskavo, alarm,** process, **ovadbo, tožbo, stečaj, prijavo, revizijo** - [METAPHORICAL]: **plaz,** mehanizem, polemiko, reakcijo, **kepo,** pobudo, akcijo, **iniciativo, aktivnost, debato, kampanjo**
kriza	1,114 (87.79 per million) - [GEOGRAPHY]: jugoslovanska, zalivska kriza - [POLITICS]: vladna, gospodarska, parlamentarna, ekonomska, ustavna, politična kriza - [METAPHORICAL]: duševna, socialna, razvojna, družbena kriza - [MODIFIERS]: huda, moralna, globoka, katastrofalna, velika, težka kriza - [NOUNS]: reševanje, razrešitev, rešitev, razplet, razreševanje krize - [VERBS]: prebroditi, poglabljati, razrešiti, povzročiti, rešiti, začeti krizo	8,062 (196.69 per million) - [GEOGRAPHY]: ukrajinska, grška, svetovna, globalna kriza - [POLITICS]: **migrantska, begunska,** gospodarska, finančna, **migracijska, humanitarna,** ekonomska, **dolžniška,** bančna, politična, **begunsko-migrantska, mlečna, javnofinančna, varnostna,** kapitalistična kriza - [METAPHORICAL]: **socialna** kriza - [MODIFIERS]: huda, **kompleksna,** globoka, velika kriza - [NOUNS]: **začetek**, breme, **izbruh, nastop, posledica, nastanek,** reševanje, **obdobje** krize - [VERBS]: kriza **nastopi, nastane, pokaže, udari** // povzročiti, reševati, poglabljati krizo

As can be seen from Table 15, the biggest difference in relative frequency between the two corpora is observed for the noun *crisis*, which is more than twice as frequent in Parlameter compared to SlovParl, despite the fact that the early 1990s were marked by a long and bloody war in the Balkans as well as severe economic hardship related to change of the economic and political system. Parlameter contains the largest number of new collocation candidates that indicate issues that were not present in the period of SlovParl, such as *migrant/refugee/humanitarian/security crisis*. On the other hand,

the secession period was marked by *constitutional/parliamentary crisis*, which are not observed in the late 2010s. Interestingly, SlovParl contains more metaphorical collocations which are not prominent in the Parlameter corpus, such as *mental/social/welfare/moral crisis*. Collocations containing geographical terms indicate the key political, military and social hotspots from that period: *Yugoslav/Gulf crisis* in early 1990s, and *Ukraine/Greek crisis* in late 2010s. An analysis of key verbal collocates with the noun crisis reveals another interesting observation, which is that in SlovParl, all the verbs are about solving the crisis (*to solve/resolve/untangle the crisis*), whereas in Parlameter, politicians mostly use verbs that discuss the beginnings or deepening of the crisis (*crisis sets in/appears/starts/hits*, *to trigger/deepen the crisis*).

The verb *trigger* is the only one of the three examples that has a higher relative frequency in SlovParl but despite the greater relative frequency, Parlameter contains more collocation candidates, both in the direct and the metaphorical sense, such as *trigger an investigation/indictment/lawsuit*, or *trigger an audit/bankruptcy*.

It is interesting to note that the adjective *southern* is more frequently used and has more collocations in general in ParlaMeter despite the fact that in the secession period, links to the rest of former Yugoslavia were probably stronger and there were probably more open issues, signalling that certain topics were probably not discussed on purpose until the issues were resolved and the relations were established again. Especially interesting are all the neighbour-related collocations, which only appear in the Parlameter corpus, 30 years after Slovenia left Yugoslavia: *southern neighbour / neighbours / neighbourhood / market / fruit,* despite the fact that geographically speaking, the former Yugoslav republics, spread south-east, not south of Slovenia. The one major unsettled issue is the border with Croatia that has even been subject of international arbitration during the parliamentary term included in the Parlameter corpus, which is reflected in the top-ranking strong collocation *južna meja/southern border*.

Conclusions

In this paper we presented the Parlameter corpus of contemporary Slovene parliamentary proceedings. We analysed the linguistic production of the speakers according to the morphosyntactic annotation of the corpus and the speaker metadata.

We have shown that despite the fact that the material included in the corpus spans the period 2014–2018, the bulk of the material was recorded in the first two full years of the parliament. When contrasted against general Slovene, parliamentary speeches contain more present tense forms and personal and demonstrative pronouns. A comparison of male and female speakers shows that while male speakers take the floor more often than their female colleagues, it is the female speakers who make longer contributions. Female speakers mostly address the topics of *health, labour, family and social affairs, public administration,* and *education, science and sport*, while most of the keywords from male speakers do not belong to specific topics, which indicate a more

discursive, debating style of the male speakers. When comparing speeches according to party lines, the most prolific deputy group is the largest opposition party Slovenian Democratic Party (SDS) while the ruling Party of Modern Centre (SMC) is the least prolific one. The most productive parties with a relative token to speaker ratio are the smallest parties in this parliamentary term, the Left (Levica) and New Slovenia (NSi). The largest opposition party SDS stands out from the rest by the large amount of ideological keywords while Levica stands out by signature stylistic devices which range from very informal to highly elevated. NSi and Levica, the opposition parties with the same number of MPs but from the opposite ends of the political spectrum, both address the widest variety of issues. With keywords belonging almost exclusively to the semantic field of retirement and pension, DeSUS lies on the other end of the spectrum as a single-issue party. A comparison with the SlovParl corpus of parliamentary debates from the period of Slovenia's independence, many more topics are identified in Parlameter, which understandable as a well-established state will need to take care of a full spectrum of issues whereas a new state will mostly be dealing with procedural issues and the new legislature. In the future we plan to enrich the corpus with additional session records of previous and the most recent parliamentary terms as well as with additional metadata available through the Parlameter system, such as voting data and accepted legislation, which are also valuable for addressing a number of research questions in various research communities. In parallel, we also plan to develop comparable corpora from other parliaments, starting with Croatian and Bosnian.

Acknowledgments

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project "Resources, methods, and tools for the understanding, identification, and classification of various forms of socially unacceptable discourse in the information society" (J7-8280, 2017–2019) and the Slovenian research infrastructure for language resources and technology CLARIN.SI.

Sources and Literature

Literature:

- Bayley, Paul. 2014. "Introduction: The Whys and Wherefores of Analyzing Parliamentary Discourse." In *Cross-Cultural Perspectives on Parliamentary Discourse*, edited by Paul Bayley, 1–44. Amsterdam, Philadelphia: John Benjamins Publishing.
- Cheng, Jennifer E. 2015. "Islamophobia, Muslimophobia or Racism? Parliamentary discourses on Islam and Muslims in Debates on the Minaret Ban in Switzerland." *Discourse & Society* 26 (5): 562–86.

- Chester, Daniel Norman, and Nona Bowring. 1962. *Questions in Parliament.* Oxford: Clarendon Press.
- van Dijk, Teun A. 2010. "Political Identities in Parliamentary Debates." In *European Parliaments Under Scrutiny: Discourse Strategies and Interaction Practices*, edited by Cornelia Ilie, 29–56. Amsterdam, Philadelphia: John Benjamins Publishing.
- Fišer, Darja, and Jakob Lenardič. 2018. "Parliamentary Corpora in the CLARIN Infrastructure." In *Selected Papers from the CLARIN Annual Conference 2017*, edited by Maciej Piasecki, 75–85. Accessed February 27, 2019. http://www.ep.liu.se/ecp/147/007/ecp17147007.pdf.
- Fišer, Darja, and Vojko Gorjanc. 2013. *Korpusna analiza*. Ljubljana: Znanstvena založba Filozofske Fakultete.
- Fišer, Darja, Nikola Ljubešić, and Tomaž Erjavec. 2018. "The Janes Project: Language Resources and Tools for Slovene user Generated Content." *Language Resources and Evaluation*. In press. https://doi.org/10.1007/s10579-018-9425-z.
- Franklin, Mark N., and Philip Norton. 1993. *Parliamentary Questions: For the Study of Parliament Group*. Oxford: Oxford University Press.
- Hirst, Graeme, Vanessa Wei Feng, Christopher Cochrane, and Nona Naderi. 2014. "Argumentation, Ideology, and Issue Framing in Parliamentary Discourse." In *ArgNLP*. Accessed 27 February 2019. ftp://www.cs.toronto.edu/pub/gh/Hirst-etal-Bertinoro-2014.pdf.
- Hughes, Lorna M., Paul S. Ell, Gareth A.G. Knight, and Milena Dobreva. 2013. "Assessing and Measuring Impact of a Digital Collection in the Humanities: An Analysis of the SPHERE (Stormont Parliamentary Hansards: Embedded in Research and Education) Project." *Digital Scholarship in the Humanities* 30 (2): 183–98.
- Ihalainen, Pasi, Cornelia Ilie, and Kari Palonen. 2016. *Parliament and Parliamentarism: A Comparative History of a European Concept.* Oxford, New York: Berghahn Books.
- Ilie, Cornelia. 2017. "Parliamentary Debates." In *The Routledge Handbook of Language and Politics*, edited by Ruth Wodak and Bernhard Forchtner. Routledge.
- Ljubešić, Nikola, and Tomaž Erjavec. 2016. "Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 1527–31. Accessed February 27, 2019. http://www.lrec-conf.org/proceedings/lrec2016/pdf/811_Paper.pdf.
- Ljubešić, Nikola, Tomaž Erjavec, Darja Fišer, Tanja Samardžić, Maja Miličević, Filip Klubička, and Filip Petkovski. 2016. "Easily Accessible Language Technologies for Slovene, Croatian and Serbian." In *Proceedings of the Conference on Language Technologies and Digital Humanities 2016*, edited by Tomaž Erjavec and Darja Fišer, 120–24. Accessed February 27, 2019. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Ljubesic-et-al_Easily-Accessible-Language-Technologies.pdf.
- Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, and Filip Dobranić. 2018. "The Parlameter corpus of contemporary Slovene parliamentary proceedings." In *Proceedings of the Conference on Language Technologies and Digital Humanities 2018,* edited by Darja Fišer and Andrej Pančur, 162–167. Accessed June 12, 2019. http://www.sdjt.si/wp/wp-content/uploads/2018/09/JTDH-2018_Ljubesic-et-al_The-Parlameter-corpus-of-contemporary-Slovene-parliamentary-proceedings.pdf
- Pančur, Andrej, and Mojca Šorn. 2016. "Smart Big Data: Use of Slovenian Parliamentary Papers in Digital History." *Prispevki za novejšo zgodovino* 56 (3): 130–46.
- Pančur, Andrej. 2016. "Označevanje zbirke zapisnikov sej slovenskega parlamenta s smernicami TEI." In *Proceedings of the Conference on Language Technologies and Digital Humanities 2016*, edited by Tomaž Erjavec and Darja Fišer, 142–48. Accessed February 27, 2019. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Pancur_Oznacevanje-zbirke-zapisnikov-sej-slovenskega-parlamenta.pdf.

- Rheault, Ludovic, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. 2016. "Measuring Emotion in Parliamentary Debates with Automated Textual Analysis." *PLoS ONE* 11 (12): 1–18.
- TEI Consortium, 2017. Guidelines for Electronic Text Encoding and Interchange. Accessed February 27, 2019. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

Sources:

- Dobranić, Filip, Nikola Ljubešić, and Tomaž, Erjavec. 2019. *Slovenian Parliamentary Corpus ParlaMeter-sl 1.0*, Slovenian Language Resource Repository CLARIN.SI. http://hdl.handle.net/11356/1208.
- Pančur, Andrej, Mojca Šorn, and Tomaž Erjavec. 2017. *Slovenian Parliamentary Corpus SlovParl 2.0*, Slovenian Language Resource Repository CLARIN.SI. http://hdl.handle.net/11356/1167.

Darja Fišer, Nikola Ljubešić, Tomaž Erjavec

PARLAMETER – A CORPUS OF CONTEMPORARY SLOVENE PARLIAMENTARY PROCEEDINGS

SUMMARY

The unique content, structure and language, as well as the availability of records of parliamentary debates are all factors that make them an important object of study in a wide range disciplines in digital humanities and social sciences. This has motivated a number of national as well as international initiatives to compile, process and analyse parliamentary corpora. This paper presents the Parlameter corpus of contemporary Slovene parliamentary proceedings, which covers the VIIth mandate of the Slovene Parliament (2014–2018). The Parlameter corpus offers rich speaker metadata (gender, age, education, party affiliation) and is linguistically annotated (lemmatization, tagging, named entity recognition).

The Parlameter corpus contains 371 sessions and 1,981 speakers who gave 133,287 speeches which contain almost 35 million words. In the paper we demonstrate the potential of the corpus analysis techniques for investigating political debates by analysing the linguistic production of the speakers according to the morphosyntactic annotation of the corpus and the speaker metadata. When contrasted against general Slovene, parliamentary speeches contain more present tense forms and personal and demonstrative pronouns. While male speakers take the floor more often than their female colleagues, the female speakers' contributions tend to be longer. Female speakers mostly address the topics of health, labour, family and social affairs, public administration, and education, science and sport, while most of the keywords from male speakers do not belong to specific topics, which indicate a more discursive, debating style of the male speakers. The most prolific deputy group overall is

the largest opposition party Slovenian Democratic Party (SDS) while the then ruling Party of Modern Centre (SMC) is the least prolific. The most productive parties with a relative token to speaker ratio are the smallest parties in that parliamentary term, the Left (Levica) and New Slovenia (NSi). The largest opposition party SDS stands out from the rest by the large amount of ideological keywords while Levica stands out by signature stylistic devices which range from very informal to highly elevated. NSi and Levica, the opposition parties with the same number of MPs but from the opposite ends of the political spectrum, both address the widest variety of issues. With keywords belonging almost exclusively to the semantic field of retirement and pension, DeSUS lies on the other end of the spectrum as a single-issue party. A comparison with the SlovParl corpus of parliamentary debates from the period of Slovenia's independence, many more topics are identified in Parlameter, which understandable as a well-established state will need to take care of a full spectrum of issues whereas a new state will mostly be dealing with procedural issues and the new legislature.

The Parlameter corpus is available through both CLARIN.SI concordancers, as well as for download from its repository, both as a TEI document and in the simpler vertical file format, under the liberal Creative Commons – Attribution-ShareAlike (CC BY-SA 4.0) licence. The corpus architecture allows for regular extensions of the corpus with additional Slovene data, as well as data from other parliaments, starting with Croatian and Bosnian.

Darja Fišer, Nikola Ljubešić, Tomaž Erjavec

PARLAMETER – KORPUS RAZPRAV SLOVENSKEGA DRŽAVNEGA ZBORA

POVZETEK

Edinstvena vsebina, struktura in jezik, pa tudi dostopnost prepisov parlamentarnih razprav so dejavniki, zaradi katerih so le-ti pomemben predmet raziskav v številnih znanstvenih disciplinah digitalne humanistike in družboslovja. To je motiviralo številne nacionalne in mednarodne iniciative za izgradnjo, označevanje in analizo parlamentarnih korpusov. V tem prispevku predstavimo korpus sodobnih parlamentarnih razprav Parlameter, ki vsebuje razprave 7. mandata slovenskega Državnega zbora (2014–2018). Korpus Parlameter vsebuje bogate metapodatke o govorcih (spol, starost, izobrazba, strankarska pripadnost) in je jezikoslovno označen (lematizacija, tegiranje, imenske entitete).

Korpus Parlameter vsebuje 371 razprav in 1.981 govorcev, ki so prispevali 133.287 govorov oziroma 35 milijonov besed. V prispevku prikažemo potencial korpusno-analitičnih tehnik za raziskovanje političnih razprav z analizo jezikovne produkcije

govorcev glede na morfosintaktične oznake in metapodatke o govorcih. Primerjava s splošno slovenščino pokaže, da v parlamentarnih govorih izstopajo sedanjiške oblike ter osebni in kazalni zaimki. Čeprav moški govorci spregovorijo večkrat, so govori žensk daljši. Ženske večinoma razpravljajo o temah, kot so zdravje, delo, družina in sociala, javna uprava ter izobraževanje, znanost in šoprt, večina ključnih besed v moških govorih pa ni vezanih na določeno tematiko, kar nakazuje bolj diskurziven, razpravljalski slog moških govorcev. V celoti gledano je najbolj produktivna strankarska skupina največja opozicijska stranka SDS, medtem ko je vladajoča stranka SMC v korpusu zastopana z najmanj izrečenimi besedami. Najvišji relativni delež števila pojavnic na govorca imata najmanjši parlamentarni stranki tega sklica Levica in NSi. Največja opozicijska stranka SDS izstopa po izrazito velikem obsegu ideološko obarvanih ključnih besed, Levica pa po specifičnih slogovnih figurah, ki so tako zelo neformalne kot zelo povzdignjene. NSi in Levica, opozicijski stranki z enakim številom poslancev a s povsem različnih polov političnega spektra, obe naslavljajta največje število tematik. Po drugi strani pa s ključnimi besedami, ki skoraj v celoti spadajo v pomensko polje upokojevanja in pokojnin, pa je povsem obratno pri stranki DeSUS, ki s tem utrjuje svoj status problemske stranke. Primerjava s korpusom SlovParl iz obdobja slovenske osamosvojitve kaže, da je v korpusu Parlameter obravnavanih veliko več tem kot v korpusu SlovParl, kar je razumljivo, saj se mora uveljavljena država ukvarjati s celotnim spektrom problematik, medtem ko se novo ustanovljena država posveča predvsem priceduralnim vprašanjem in sprejemanju nove zakonodaje.

Korpus Parlameter je dostopen preko obeh konkordančnikov v okviru raziskovalne infrastructure CLARIN.SI, prav tako pa ga je mogoče prenesti z repozitorija v format TEI, pa tudi v preprostejšem vertikalnem formatu pod licenco Creative Commons – Attribution-ShareAlike (CC BY-SA 4.0). Korpusna arhitektura je zasnovana tako, da omogoča širitev korpusa na druga časovna obdobja, prav tako pa tudi vključevanje gradiv drugih parlamentov, začenši s hrvaškim in bosanskim.