MLO Modern Languages Open

Fišer, D and Pahor de Maiti, K 2020 Voices of the Parliament. *Modern Languages Open*, 2020(1): 46 pp. 1–23. DOI: https://doi.org/10.3828/mlo.v0i0.295

ARTICLE – DIGITAL MODERN LANGUAGES

Voices of the Parliament

Darja Fišer and Kristina Pahor de Maiti

University of Ljubljana, Faculty of Arts, SI

Corresponding authors: Darja Fišer (darja.fiser@ff.uni-lj.si); Kristina Pahor de Maiti (kristina.pahordemaiti@ff.uni-lj.si)

This tutorial shows how corpora can be used to investigate language use and communication practices in a specialised socio-cultural context of political discourse. We will demonstrate the potential of a richly annotated diachronic corpus of Slovenian parliamentary debates for investigating the characteristics and dynamics of the representation of women and their interests in the parliament.

1 Introduction

Parliaments represent the main fora for political debate that shapes legislation which directly impacts everyday life. Parliamentary discourse is motivated by a wide range of communicative roles and reveals patterns of political agendas, ideological stances and institutional roles of members of parliament who represent the interests of the citizens of a country (Ilie). This is also why parliaments have always been of interest to scholars from a range of disciplines in the humanities and social sciences.

Parliamentary proceedings are increasingly being made available in a digitised form and have been turned into structured linguistic resources called corpora for many European languages. These are often available online and can be queried through dedicated tools called concordancers. Researchers use them to perform diverse linguistic, stylistic, cultural, societal and political studies (Biel et al.; Jaworska & Ryan).

While corpus methods are widely used in linguistics (McEnery & Hardie; Biber & Reppen), including gender analysis (Baker), this tutorial shows the potential of richly annotated language corpora for research of the socio-cultural context and changes over time that are reflected through language use. The tutorial encourages students and scholars of modern languages, but also users from other fields of digital humanities and social sciences who are interested in the study of socio-cultural phenomena through language, to engage with user-friendly digital tools for the analysis of large text collections. The tutorial is designed in a way that takes full advantage of both linguistic annotations and the available speaker and text metadata to formulate powerful quantitative queries that are then further extended with manual qualitative analysis in order to ensure adequate framing and interpretation of the results.

The tutorial demonstrates the potential of parliamentary corpora research via concordancers without the need for programming skills. No prior experience in using language corpora and corpus querying tools is required in order to follow this tutorial. While the same analysis could be carried out on any parliamentary corpus with similar annotations and metadata, in this tutorial we will use the siParl corpus which contains parliamentary debates of the National Assembly of the Republic of Slovenia from 1990 to 2018, but no knowledge of Slovenian is required to follow the tutorial. To try out the analyses in other languages, we invite you to explore a parliamentary corpus of your choice from those available through CLARIN.

2 Practicalities

This tutorial starts with a brief introduction to corpora and corpus analysis, followed by an introduction on the characteristics of specialised corpora of parliamentary debates and an overview of research into language and gender. The second part of the tutorial is a hands-on section which demonstrates the potential of some of the best-known corpus analysis techniques, such as concordances, frequency lists, keywords and collocations, to explore the topics female speakers debate in parliament over time and to compare and contrast their language use with the language of their male counterparts.

All the resources and tools used in this tutorial are online and available under open license. Corpus querying will be demonstrated on the noSketch Engine concordancer, while additional manual analysis and visualisation of the results will be performed in a spreadsheet editor (e.g. Google Spreadsheet or MS Excel).

Screencasts, explanations of corpus-querying procedures and links to the results are provided in green boxes for anyone who wishes to reproduce the searches on their own.

The siParl corpus can be queried online through the noSketch Engine concordancers at CLARIN.SI, the Slovenian node of CLARIN ERIC, the European research infrastructure for language resources and technology. siParl can also be downloaded from the CLARIN.SI repository and then further analysed with other corpus or text mining tools. Tutorials showing how this can be done are available online (e.g. Corpus Analysis with Antconc and Basic Text Processing in R).

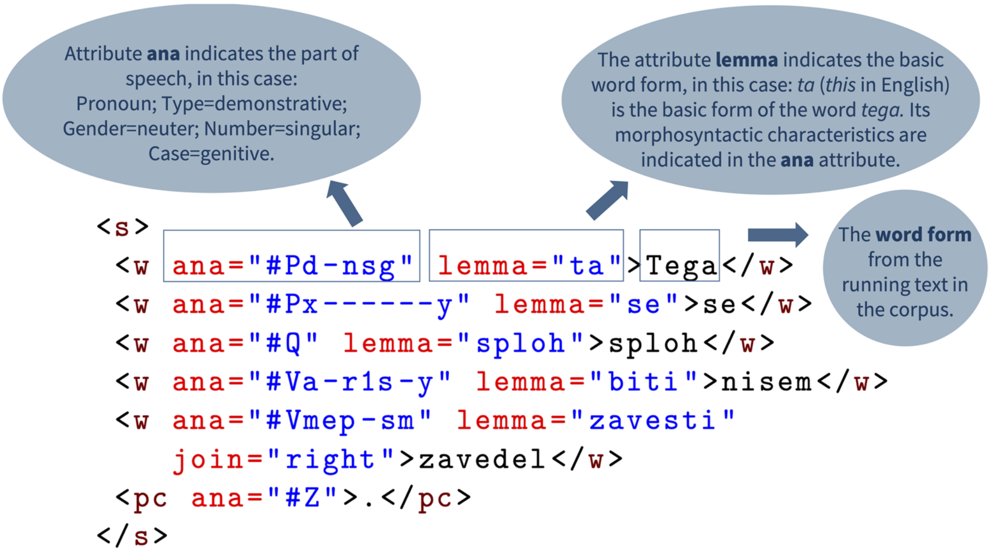
3 Corpora and concordancers

3.1 Corpora

Language corpora are large collections of carefully selected machine-readable language data which can be used as the basis of linguistic descriptions or as a means of verifying hypotheses about a language. However, corpora are much more than just simple collections of texts in electronic form. They are formatted in one of the standard formats, such as the Extensible Markup Language or XML, and encoded according to a predetermined but usually flexible schema for the representation of texts in digital form. One of the most established encoding methods in linguistics and digital humanities is the Text Encoding Initiative or TEI.

The basic unit in the corpus is the token, which is a sequence of characters separated by white space, such as a word form, number or punctuation. To facilitate corpus search, texts in corpora are linguistically annotated. The two most basic forms of linguistic annotations are part of speech tagging, which marks up running words in texts with their part of speech, and lemmatisation, which is the assignment of base forms or lemmas to tokens (word forms) that are especially important for corpora of morphologically rich languages, such as Slovenian. In addition to linguistic annotation, corpora are also enriched with text and speaker metadata that are needed for the contextualisation of the search results but can also be used for more fine-grained corpus querying.

Figure 1 shows an excerpt from the corpus of Slovenian parliamentary debates which has been encoded in TEI (the <*s*> element marks up sentences, <*w*> words and <*pc*> punctuation) and annotated with part of speech tags (given in the *ana* attribute) and lemmas (given in the *lemma* attribute).



Attribute **ana** indicates the part of speech, in this case: Pronoun; Type=demonstrative; Gender=neuter; Number=singular; Case=genitive.

The attribute **lemma** indicates the basic word form, in this case: *ta* (*this* in English) is the basic form of the word *tega*. Its morphosyntactic characteristics are indicated in the **ana** attribute.

The **word form** from the running text in the corpus.

```
<s>
 <w ana="#Pd-nsg" lemma="ta">Tega</w>
 <w ana="#Px------y" lemma="se">se</w>
 <w ana="#Q" lemma="sploh">sploh</w>
 <w ana="#Va-r1s-y" lemma="biti">nisem</w>
 <w ana="#Vmep-sm" lemma="zavesti"
    join="right">zavedel</w>
 <pc ana="#Z">.</pc>
</s>
```

Figure 1: Excerpt from a linguistically annotated and TEI encoded corpus.

3.2 Concordancers

Corpora are queried with specialised corpus analysis tools, also called concordancers. Concordancers are either installed on a computer or accessed through a website and can be used to retrieve all instances of a sequence of tokens from the corpus. Many different concordances with similar functionalities exist (see this detailed list) but some of the most popular ones are the offline AntConc (free) and WordSmith Tools (licence needs to be purchased) where it is necessary to load and query your own corpora, and the online BYU concordancer (free registration required for full functionality) and SketchEngine (free licence for students, teachers and researchers from academic institutions in the EU) which offer a lot of extensive preloaded corpora for many languages as well as the possibility to build and annotate your own.

Most of the modern concordancing tools offer the following basic corpus analysis techniques that will also be used in this tutorial:

- Concordances show all the hits of the searched word or phrase in the corpus together with its context. They can be randomised or sorted according to the searched word or phrase or by its left or right context, revealing typical patterns in which it is used.
- Word lists summarise the frequencies of all the hits in the corpus that correspond to the corpus query out of context and can be sorted alphabetically or by frequency.
- Keyword lists highlight what words are prominent in a focus corpus compared to a reference one.
- Collocation lists return words that are typically combined with the searched word.

Dedicated tutorials for concordancers (e.g. the SketchEngine Quick Start Guide) as well as general corpus-linguistic courses (e.g. the course Corpus Linguistics: Method, Analysis, Interpretation) are already available online, and we will refer to them wherever possible. The focus of this tutorial is to showcase how the functionalities of a popular concordance tool can be utilised and combined on a specialised corpus of parliamentary data in order to answer several real-world research questions in a methodologically sound way, as such skills are still lacking, especially for students and scholars of modern languages and other fields of (digital) humanities and social sciences who are interested in the study of socio-cultural phenomena through language use.

4 Parliamentary records

Parliamentary corpora are of different sizes and contain data of different modalities (written, spoken, gestural) and time periods in one or multiple languages. Although parliamentary debates are mainly provided as written texts (e.g. The Danish Parliament Corpus), they are sometimes also accessible as audio/video recordings coupled with corresponding transcriptions (e.g. Czech Parliamentary Meetings). When used as a diachronic source, a parliamentary corpus enables in-depth research on linguistic and societal change over time. Most parliamentary corpora have rich metadata about the speeches (e.g. date of the speech, duration of the speech, agenda item to which the speech is related) or the speakers (e.g. gender, age, education, political affiliation, institutional role) which offer valuable insights into the context on the studied phenomenon (see Alasuutari et al.; Demmen et al.).

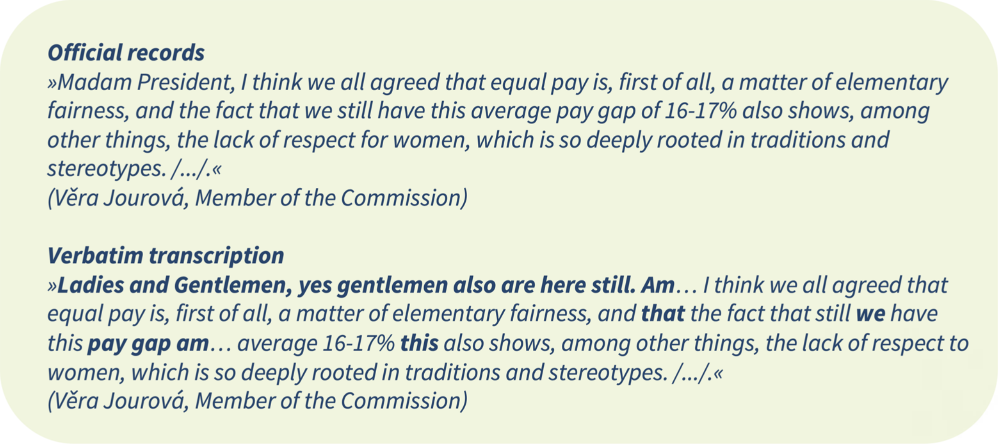
4.1 Parliamentary discourse

Parliaments are institutions governed by specific rules and conventions. These are shaped by socio-historical traditions that influence the organisation and operation of the parliament which also extend to language use, such as conventions for turn-taking or forms of address. These conventions, however, are not necessarily shared among different parliaments; for instance, interruptions are strictly prohibited in the Greek parliament, whereas in the UK parliament these are common and often not sanctioned (Ilie), which needs to be taken into consideration when forming queries or interpreting the results. Therefore, whenever this type of discourse is investigated, it is particularly important to understand the circumstances it was produced in.

4.2 Faithfulness of the records

It should be noted that officially released records of parliamentary debates are not 100% verbatim transcriptions and that minute-taking practices vary through history and across countries. Editing usually involves elimination of some typical characteristics of spoken language but may also include other interventions, such as the elimination of obvious language or factual errors, dialectal or colloquial expressions, and rude and obscene language. Editing guidelines are mostly not made public, which can substantially hinder research (for more source criticism, see Mollin; Rix). Furthermore, speeches by MPs are often prepared in advance (i.e. written-to-be-spoken) which has a big impact on their stylistic features. These peculiarities together with the broader socio-historical context always need to be taken into account when defining research questions, methodology and the interpretation of the results.

Figure 2 shows an illustrative comparison of the official records of a speech by Věra Jourová, a Czech politician and European Commissioner for Justice, Consumers and Gender Equality (2014–2019) at the plenary session debate on the gender pay gap from 1 May 2017 published on the website of the EU Parliament, together with a verbatim transcription of the video clip of her speech created for this tutorial. A quick look at the differences in the two transcripts shows that transcriptions of EU Parliamentary debates might not be the best resource for studying the use of determiners, frequency of hesitations or false starts or even the use of spontaneous humour in parliamentary speech.



Official records

»Madam President, I think we all agreed that equal pay is, first of all, a matter of elementary fairness, and the fact that we still have this average pay gap of 16-17% also shows, among other things, the lack of respect for women, which is so deeply rooted in traditions and stereotypes. /.../.«

(Věra Jourová, Member of the Commission)

Verbatim transcription

*»**Ladies and Gentlemen, yes gentlemen also are here still. Am**… I think we all agreed that equal pay is, first of all, a matter of elementary fairness, and **that** the fact that still **we** have this **pay gap am**… average 16-17% **this** also shows, among other things, the lack of respect to women, which is so deeply rooted in traditions and stereotypes. /.../.«*

(Věra Jourová, Member of the Commission)

Figure 2: Comparison of official records and verbatim transcriptions from the EU Parliament.

4.3 Know your research dataset

In order to make best use of any given corpus, to formulate search queries correctly and interpret the results appropriately, it is important to understand what the selected corpus contains, how it was constructed and annotated, and what its limitations are. The level of annotation varies from corpus to corpus, so the details for the selected corpus should always be checked. Such information is typically found on a dedicated webpage (e.g. Corpus of Historical Low German), inside the concordancer through which it is made available (e.g. Hansard in the concordancer of Brigham Young University) or in the repository where the corpus is archived (e.g. ParlaMeter-hr in the CLARIN.SI repository). You can read more about the encoding and annotation of parliamentary data in the Parthenos training module on the collections of parliamentary records.

5 Language and gender

Differences between the language of men and women have been the focus of much research in sociolinguistics, stylistics, rhetoric, gender studies, media studies and discourse analysis (see Eckert & McConnell-Ginet; Wodak). The results show that the differences are subtle but systematic (Newman et al.). Yet political communication studies have traditionally been based on male politicians; only recently have scholars begun to consider the discourse of female politicians (see Marshall).

This is important because several authors (Antić Gaber; Leijenaar; Wolbrecht) have demonstrated that female legislators differ from their male counterparts in the issues they address, the positions they take, and the approach they use in law-making. In their analysis of the representation of women in the UK parliament after 1945, Blaxill and Beelen showed a stronger emphasis on the women's issues in speeches delivered by female MPs, who on average also contributed considerably more speeches about women in comparison to male MPs. In political science (cf. Osborn), women's issues are considered those issues that are, on one hand, traditionally believed to be in the domain of women (for example, education, healthcare etc.), and on the other, that concern women's wellbeing directly (for example, child care, domestic violence, equality etc.). Similarly, Bäck et al., Hansen et al., Mensah and Wood found in the corpora of parliamentary speeches from Sweden, Denmark and Ghana, respectively, that women more often spoke about the soft policy areas in comparison to men, who more often tackle the hard policy areas. The terminology is adopted from political science (e.g. Wängnerud), where policy areas are divided into so-called hard (e.g. macroeconomics, energy, transport, banking, finance and domestic commerce, space, science and technology, and communications) and soft ones (e.g. health, labour, employment and immigration, education, and social welfare).

We should note, however, that various parameters (e.g. social class, context, age, hierarchy) have been shown to influence language use and that gender is only one of them (Coates; Litosseliti). Furthermore, as Bing and Bergvall (quoted in Litosseliti) point out, similarities in language use of different genders are often overlooked, despite being more significant than the differences. Likewise, Blaxill and Beelen have shown a similar tendency in the context of parliamentary discourse research. So, we always need to be careful not to jump to quick conclusions and over-interpret the results for the features we expect to see because we know the gender of the speaker (Goddard & Patterson).

6 Corpus analysis

In this section, we will explore a large corpus of Slovenian parliamentary debates. We will demonstrate how basic corpus analysis techniques can be used to answer three different research questions:

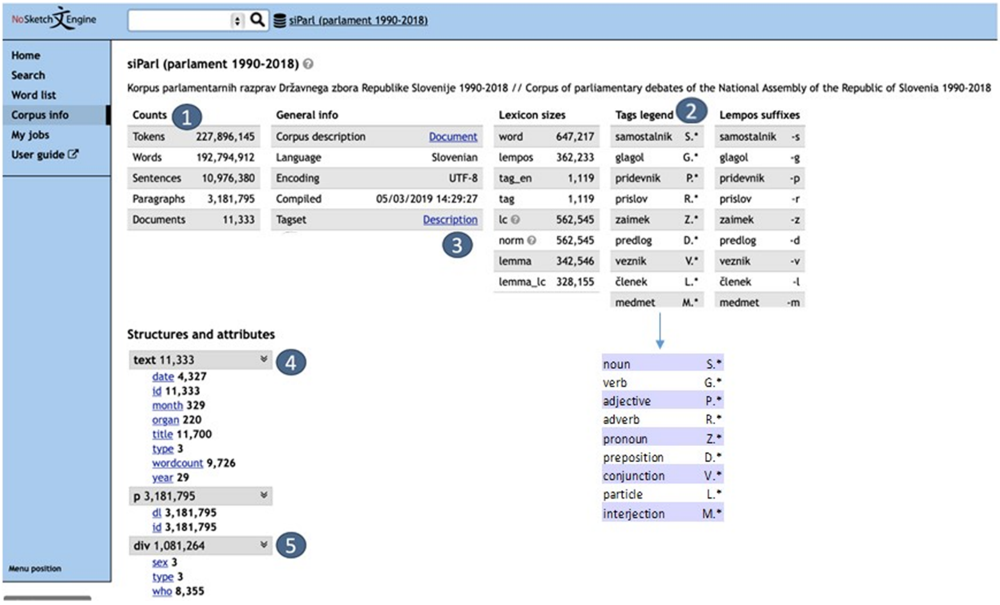
- In Task 1, we will analyse the representation of women in the Slovenian parliament. To do this, we will first learn **how to create subcorpora**. Then, we will learn **how to build frequency lists** showing the number of speakers and their contribution in the subcorpora.
- In Task 2, we will examine the most prominent topics discussed by female speakers in comparison to their male counterparts and over time. We will gain an insight into the topics by first learning **how to extract keywords** from the subcorpora. Next, in order to analyse their usage in context, we will learn **how to perform concordance analysis**.
- In Task 3, we will investigate how women's issues have been debated by female and male speakers since the first fully democratic elections in independent Slovenia in 1992. We will first learn **how to use and compare relative frequencies** in subcorpora of different sizes. Then, we will see **how to extract collocations** of selected nouns in the subcorpora.

6.1 The siParl corpus

The siParl corpus (Pančur et al.) is composed of parliamentary debates of the National Assembly of the Republic of Slovenia from 1990 to 2018. The corpus contains records of parliamentary debates from the period of secession from Yugoslavia until the end of the seventh parliamentary term, minutes of the working bodies of the National Assembly of the Republic of Slovenia from the second to the seventh parliamentary term 1996–2018, and minutes of the Council of the President of the National Assembly also from the second to the seventh parliamentary term.

The corpus contains metadata about the speakers, a typology of parliamentary sessions and structural and editorial annotations. It is also linguistically annotated for parts of speech and lemma. The corpus comprises over a million speeches or 195 million words delivered by more than 8,000 speakers (e.g. members of parliament, members of the government, ministry representatives, representatives of professional organisations, non-governmental organisations and interest groups).

This information is also summarised in the concordancer (see **Figure 3**):



siParl (parlament 1990-2018)

Korpus parlamentarnih razprav Državnega zbora Republike Slovenije 1990-2018 // Corpus of parliamentary debates of the National Assembly of the Republic of Slovenia 1990-2018

Counts	
Tokens	227,896,145
Words	192,794,912
Sentences	10,976,380
Paragraphs	3,181,795
Documents	11,333

General info	
Corpus description	Document
Language	Slovenian
Encoding	UTF-8
Compiled	05/03/2019 14:29:27
Tagset	Description

Lexicon sizes	
word	647,217
lempos	362,233
tag_en	1,119
tag	1,119
lc	562,545
norm	562,545
lemma	342,546
lemma_lc	328,155

Tags legend	
samostalnik	S.*
glagol	G.*
pridevnik	P.*
prislov	R.*
zaimek	Z.*
predlog	D.*
veznik	V.*
členek	L.*
medmet	M.*

Lempos suffixes	
samostalnik	-s
glagol	-g
pridevnik	-p
prislov	-r
zaimek	-z
predlog	-d
veznik	-v
členek	-l
medmet	-m

noun	S.*
verb	G.*
adjective	P.*
adverb	R.*
pronoun	Z.*
preposition	D.*
conjunction	V.*
particle	L.*
interjection	M.*

Structures and attributes

text 11,333
- date 4,327
- id 11,333
- month 329
- organ 220
- title 11,700
- type 3
- wordcount 9,726
- year 29

p 3,181,795
- dl 3,181,795
- id 3,181,795

div 1,081,264
- sex 3
- type 3
- who 8,355

Figure 3: The siParl information page.

- Quantitative information about the size of the corpus is provided in the *Counts* section (see *Item 1*).
- The basic tags for parts of speech are listed in *Tags legend* (see *Item 2*) while the full tagset description is available through the link (see *Item 3*).
- The corpus is encoded in the form of structural attributes at the *text* level which represents a single session (see *Item 4*), such as session date, session type (Upper House, Lower House, Working Body) or parliamentary organ (e.g. Constitutional Commission, Committee on European Union Affairs, Subcommittee on Roma Issues etc.).
- In addition, structural attributes at the *div* level, which represents one utterance of a speaker (see *Item 5*), give information on their gender, their role (type attribute) and their name.

6.2 TASK 1: Representation of women in the Slovenian parliament

The Slovenian National Assembly has ninety MPs, including one representative for the Italian and one for the Hungarian minority, who are members of currently nine political parties. Slovenia is one of the youngest EU states and has seen dramatic changes in gender equality since the early 1990s. While in 1990 (when Slovenia was still part of the Federal Socialist Republic of Yugoslavia) the share of women in parliament was 24%, this share dropped dramatically with the first multi-party elections. When the National Assembly of the Republic of Slovenia was convened for the first time in 1992, only a dozen (13%) female members held seats (Selišnik & Antić Gaber). During the transition process when social, political, economic and value systems fundamentally changed, women in Slovenia lost more of the economic and social gains of socialism than men and were almost completely ousted from key political institutions. But in the seventh parliamentary term (1 August 2014–22 July 2018), which is the last one included in the siParl corpus, female members of parliament held 35% of the seats, largely due to legislative measures enforcing gender quotas. According to the EU gender equality value and political power index for 2017, Slovenia ranked among the top ten EU countries in terms of the proportion of women MPs.

In Task 1, we are interested in comparing the contents of the siParl corpus with the trends observed in parliamentary elections and the Slovenian society.

6.2.1 Creating subcorpora

By taking advantage of the metadata available in the corpus (see **Figure 3**), we split the corpus into parts, called subcorpora, according to the following criteria:

- **Gender of speaker.** Each speaker in the corpus is labelled with one of the following gender categories: male, female or unknown (in cases when the metadata records are incomplete). In this tutorial we will only use the first two categories, male and female.
- **Type of speaker.** Each speaker in the corpus is labelled with one of the following speaker types: regular speaker, presiding speaker or unauthorised speaker. Regular speakers are all the speakers in the parliament who have been explicitly given the floor by the presiding speaker. In addition to the members of the parliament, these can be members of the government, representatives of the ministries, non-governmental organisations and so forth. The category of unauthorised speakers is very rare in the corpus and is assigned to speakers who have not been given the floor by the presiding speaker and are interrupting another speaker or speaking uninvited. For this tutorial we only use the category of regular speakers. We have intentionally excluded the presiding speakers because most of their speeches are regulated by bylaws and other procedures, and are not influenced by their party affiliation, gender or other factors, and would as such skew the results.
- **Parliamentary term.** Each speech in the corpus is labelled with the date of the speech, which we used to group together all the speeches made in each of the seven parliamentary terms that the corpus covers.

Based on these criteria we created a total of fourteen subcorpora, one for each of the seven parliamentary terms that contains the speeches of female and male speakers, respectively.

A screencast of how to create a subcorpus in noSketch Engine is available here.

The created subcorpora and information on their size are available here.

For advanced users of the tutorial we also provide an example of the CQL commands that were used to generate the subcorpora:

<div sex="ženski" & type="Regularni govornik"/> within <text type="Spodnji dom" & date>="1992-12-23" & date<="1996-11-28"/>

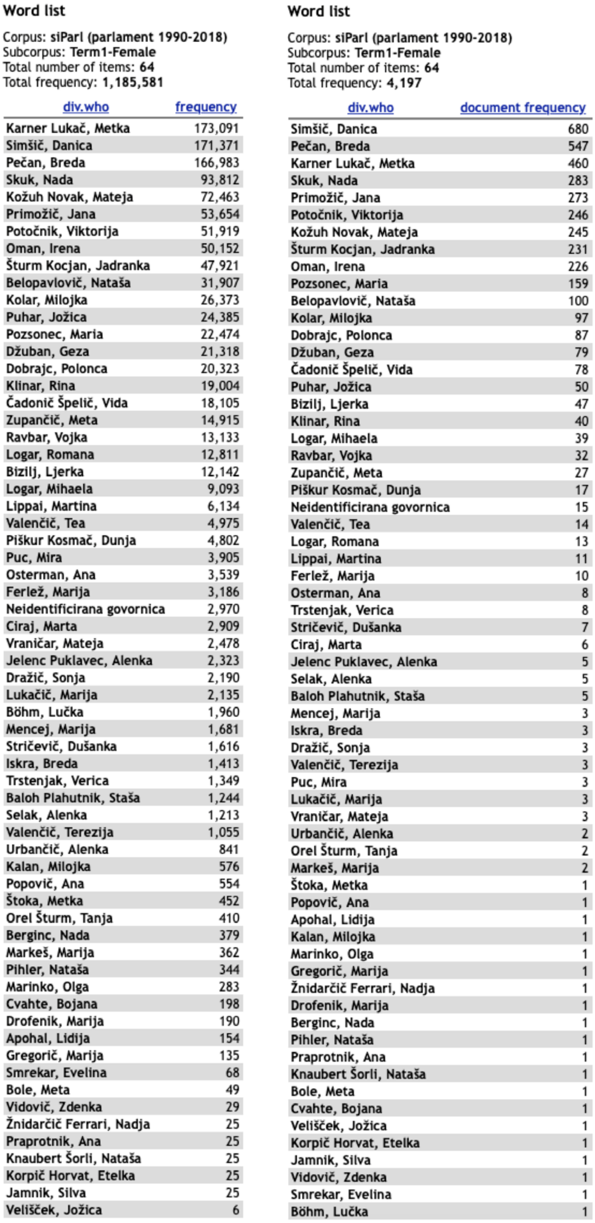
This command searches for all utterances (*div*) spoken by speakers whose gender is *female (*Slo. ženski*)* and whose type is *regular (*Slo. *Regularni govornik)* within texts from the *Lower House* (Slo. *Spodnji dom*) in the period between *23 December 1992* and *28 November 1996*, which corresponds to the first parliamentary term.

6.2.2 Using frequency lists

Taking advantage of the subcorpora created in the previous step, we will analyse the contributions of female and male speakers across time with the help of one of the most basic corpus techniques, frequency lists. These display query results from most frequent to least frequent. This technique can, for example, be used to build a frequency list of all the words uttered in the whole parliament, or by a specific speaker. In this task, we will use the function to obtain information on the number of male and female speakers in each parliamentary term, and the number of tokens they uttered.

A screencast of how to create a frequency list in noSketch Engine is available here.
An example of a frequency list can be seen in **Figure 4**.

As can be seen from **Figure 4**, there were a total of sixty-four female speakers in the first parliamentary term who collectively uttered just under 1.2 million tokens or nearly 4,200



Word list

Corpus: **siParl (parlament 1990-2018)**
Subcorpus: **Term1-Female**
Total number of items: **64**
Total frequency: **1,185,581**

div.who	frequency
Karner Lukač, Metka	173,091
Simšič, Danica	171,371
Pečan, Breda	166,983
Skuk, Nada	93,812
Kožuh Novak, Mateja	72,463
Primožič, Jana	53,654
Potočnik, Viktorija	51,919
Oman, Irena	50,152
Šturm Kocjan, Jadranka	47,921
Belopavlovič, Nataša	31,907
Kolar, Milojka	26,373
Puhar, Jožica	24,385
Pozsonec, Maria	22,474
Džuban, Geza	21,318
Dobrajc, Polonca	20,323
Klinar, Rina	19,004
Čadonič Špelič, Vida	18,105
Zupančič, Meta	14,915
Ravbar, Vojka	13,133
Logar, Romana	12,811
Bizilj, Ljerka	12,142
Logar, Mihaela	9,093
Lippai, Martina	6,134
Valenčič, Tea	4,975
Piškur Kosmač, Dunja	4,802
Puc, Mira	3,905
Osterman, Ana	3,539
Ferlež, Marija	3,186
Neidentificirana govornica	2,970
Ciraj, Marta	2,909
Vraničar, Mateja	2,478
Jelenc Puklavec, Alenka	2,323
Dražič, Sonja	2,190
Lukačič, Marija	2,135
Böhm, Lučka	1,960
Mencej, Marija	1,681
Stričevič, Dušanka	1,616
Iskra, Breda	1,413
Trstenjak, Verica	1,349
Baloh Plahutnik, Staša	1,244
Selak, Alenka	1,213
Valenčič, Terezija	1,055
Urbančič, Alenka	841
Kalan, Milojka	576
Popovič, Ana	554
Štoka, Metka	452
Orel Šturm, Tanja	410
Berginc, Nada	379
Markeš, Marija	362
Pihler, Nataša	344
Marinko, Olga	283
Cvahte, Bojana	198
Drofenik, Marija	190
Apohal, Lidija	154
Gregorič, Marija	135
Smrekar, Evelina	68
Bole, Meta	49
Vidovič, Zdenka	29
Žnidarčič Ferrari, Nadja	25
Praprotnik, Ana	25
Knaubert Šorli, Nataša	25
Korpič Horvat, Etelka	25
Jamnik, Silva	25
Veliščček, Jožica	6

Word list

Corpus: **siParl (parlament 1990-2018)**
Subcorpus: **Term1-Female**
Total number of items: **64**
Total frequency: **4,197**

div.who	document frequency
Simšič, Danica	680
Pečan, Breda	547
Karner Lukač, Metka	460
Skuk, Nada	283
Primožič, Jana	273
Potočnik, Viktorija	246
Kožuh Novak, Mateja	245
Šturm Kocjan, Jadranka	231
Oman, Irena	226
Pozsonec, Maria	159
Belopavlovič, Nataša	100
Kolar, Milojka	97
Dobrajc, Polonca	87
Džuban, Geza	79
Čadonič Špelič, Vida	78
Puhar, Jožica	50
Bizilj, Ljerka	47
Klinar, Rina	40
Logar, Mihaela	39
Ravbar, Vojka	32
Zupančič, Meta	27
Piškur Kosmač, Dunja	17
Neidentificirana govornica	15
Valenčič, Tea	14
Logar, Romana	13
Lippai, Martina	11
Ferlež, Marija	10
Osterman, Ana	8
Trstenjak, Verica	8
Stričevič, Dušanka	7
Ciraj, Marta	6
Jelenc Puklavec, Alenka	5
Selak, Alenka	5
Baloh Plahutnik, Staša	5
Mencej, Marija	3
Iskra, Breda	3
Dražič, Sonja	3
Valenčič, Terezija	3
Puc, Mira	3
Lukačič, Marija	3
Vraničar, Mateja	3
Urbančič, Alenka	2
Orel Šturm, Tanja	2
Markeš, Marija	2
Štoka, Metka	1
Popovič, Ana	1
Apohal, Lidija	1
Kalan, Milojka	1
Marinko, Olga	1
Gregorič, Marija	1
Žnidarčič Ferrari, Nadja	1
Drofenik, Marija	1
Berginc, Nada	1
Pihler, Nataša	1
Praprotnik, Ana	1
Knaubert Šorli, Nataša	1
Bole, Meta	1
Cvahte, Bojana	1
Veliščček, Jožica	1
Korpič Horvat, Etelka	1
Jamnik, Silva	1
Vidovič, Zdenka	1
Smrekar, Evelina	1
Böhm, Lučka	1

Figure 4: Frequency lists of the female speakers from the first parliamentary term. On the left, we display the number of tokens the speakers have uttered, and on the right, we show the number of speeches they have made.

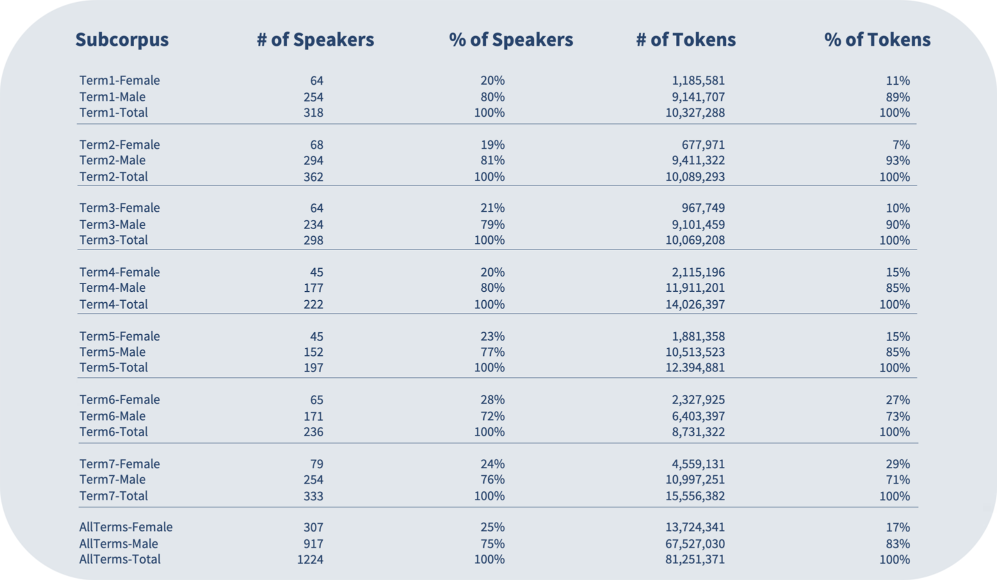
speeches. This is 18,500 tokens or sixty-five speeches per speaker on average. However, the contribution per speaker is very uneven, ranging from more than 170,000 tokens or nearly 700 speeches to as little as one speech comprising only six tokens.

Measured in tokens, the female speaker with the largest contribution in this subcorpus is Metka Karner Lukač, member of the Slovenian People's Party – the oldest party in Slovenia, who uttered 173,091 tokens or 15% of the entire subcorpus, which is nearly ten times the average. The female speaker with the smallest contribution is Jožica Velišček, the Secretary-General of the National Assembly, whose role is essential in the organisation of the work of the parliament and its working bodies but not as regular speaker in the parliament. In fact, a concordance search reveals that the six tokens actually represent a sentence she uttered as an intervention on a procedural matter.

Measured in speeches, the first-ranking speaker is Danica Simšič, member of the Democratic Party of Slovenia, a small opposition party which was elected to the parliament only in the first term, who contributed 680 speeches or 15% of all the speeches in the entire subcorpus, which is more than ten times the average. It is interesting to note that as many as twenty speakers, which is a third of all female speakers in the first parliamentary term, only spoke once. However, none of those were MPs but guest speakers. The elected MP with the lowest number of speeches as well as words is in fact Mihaela Logar, member of the Slovenian People's Party, who spoke thirty-nine times with just over 9,000 tokens in her four-year term as MP, which is nearly twenty times less than the highest ranking speakers.

6.2.3 Comparative analysis

For a comparative analysis of the representation of men and women in the Slovenian parliament over time, we recorded the number of speakers and the number of tokens they contributed for each of the fourteen subcorpora created in section Creating subcorpora. We entered them into a spreadsheet, as can be seen in **Table 1**. We also visualised the results, as can be seen in **Figures 5** and **6**.



Subcorpus	# of Speakers	% of Speakers	# of Tokens	% of Tokens
Term1-Female	64	20%	1,185,581	11%
Term1-Male	254	80%	9,141,707	89%
Term1-Total	318	100%	10,327,288	100%
Term2-Female	68	19%	677,971	7%
Term2-Male	294	81%	9,411,322	93%
Term2-Total	362	100%	10,089,293	100%
Term3-Female	64	21%	967,749	10%
Term3-Male	234	79%	9,101,459	90%
Term3-Total	298	100%	10,069,208	100%
Term4-Female	45	20%	2,115,196	15%
Term4-Male	177	80%	11,911,201	85%
Term4-Total	222	100%	14,026,397	100%
Term5-Female	45	23%	1,881,358	15%
Term5-Male	152	77%	10,513,523	85%
Term5-Total	197	100%	12.394,881	100%
Term6-Female	65	28%	2,327,925	27%
Term6-Male	171	72%	6,403,397	73%
Term6-Total	236	100%	8,731,322	100%
Term7-Female	79	24%	4,559,131	29%
Term7-Male	254	76%	10,997,251	71%
Term7-Total	333	100%	15,556,382	100%
AllTerms-Female	307	25%	13,724,341	17%
AllTerms-Male	917	75%	67,527,030	83%
AllTerms-Total	1224	100%	81,251,371	100%

Table 1: Overview of the sizes of the siParl subcorpora.

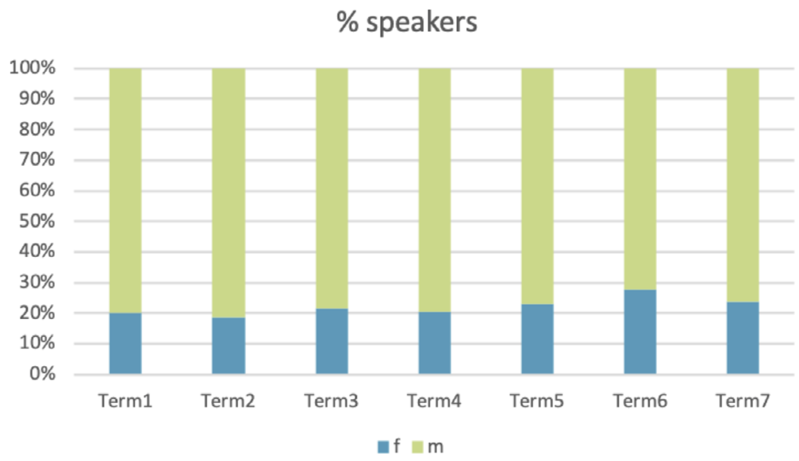


Figure 5: Share of male and female speakers in siParl over time.

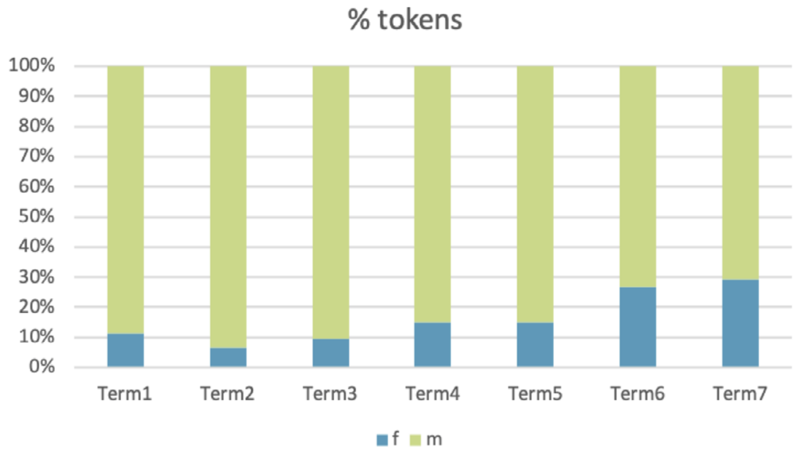


Figure 6: Contribution of male and female speakers in siParl measured in tokens uttered over time.

It is important to bear in mind that because the subcorpora were built using the *regular speaker* filter, they contain speeches produced by all female or male speakers in the parliament, not just the elected members of parliament. Nevertheless, the general trend of the representation of women in the parliament in **Figure 5** is in line with previous findings by social and political scientists, with the number of female speakers ranging from one fifth at the time of independence to about one quarter by the end of the period included in the corpus.

Importantly, however, there is a large discrepancy between the number of female speakers and the amount of words uttered by them that previous work did not account for because their authors worked with election results only. Our corpus-based approach thus complements their work with new findings which show that for female speakers, there is a clear tendency to contribute a lower share of the content than would be expected given their share of speakers. In parliamentary terms with the lowest share of female speakers, women also say the least, especially in the second parliamentary term where they produced nearly three times fewer words than would correspond to their speaker share. In recent years, as the number of female speakers in the parliament increased, their volume started increasing as well, with the seventh parliamentary term being the first time when about a quarter of female speakers produced nearly a third of all the volume.

6.3 TASK 2: Issues addressed by women

Studies of female political discourse (see section Language and gender) have shown that women tend to debate different topics to men.

In Task 2, we are interested in comparing the topics discussed by female speakers in siParl with their male counterparts. While topic classification can also be automatic (as in Karan et al.), the goal of this tutorial is to demonstrate the potential of parliamentary corpora research via concordancers without requiring programming skills, which is why we have opted for a two-step manual approach. Moreover, manual approaches are especially appropriate for highly specific topic classification tasks such as ours, for which no pre-trained models or training data exist.

6.3.1 Extracting keywords

To enable the comparative analysis we will first use a common corpus analysis technique called keyword extraction. It compares a focus corpus against a reference corpus in order to identify the most distinguishing vocabulary of the focus corpus. The focus corpus is the corpus or subcorpus under investigation. The reference corpus is typically a large representative corpus of a given language but can also be any other corpus or subcorpus we wish to use as the reference for comparison. In this task, we will contrast the female-male pairs of the siParl subcorpora to uncover the most prominently discussed issues by speakers in the Slovenian parliament.

A screencast of how to generate a keyword list in noSketch Engine is available here.

The generated keyword lists (using lowercased lemmas) are available here:

– Term1-Female

– Term1-Male

– Term7-Female

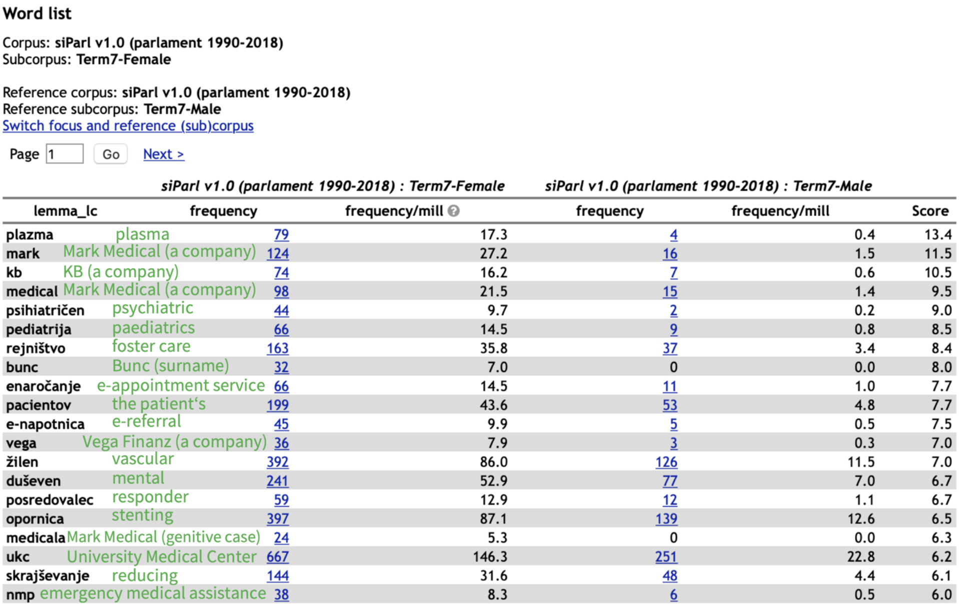
– Term7-Male

Keyword lists for each of the four subcorpora were exported into a spreadsheet and manually annotated for topics as can be seen in **Tables 3, 4** and **5**.

Figures 7 and **8** show the twenty top-ranking keywords for female and for male speakers in the last parliamentary term included in the corpus, using the Simple Maths statistics. Stark differences can be observed: while nearly all the displayed keywords for female speakers are related to healthcare, the key male vocabulary belongs to the domains of transport and foreign affairs.

6.3.2 Analysing concordances

For further analysis we selected the 100 top-ranking key lemmas, excluding person names, on the four generated keyword lists. Our goal was to manually categorise each of them into topics after inspecting their concordances, which are lists of all instances of the search word in their context as shown in **Figure 9**:



Word list

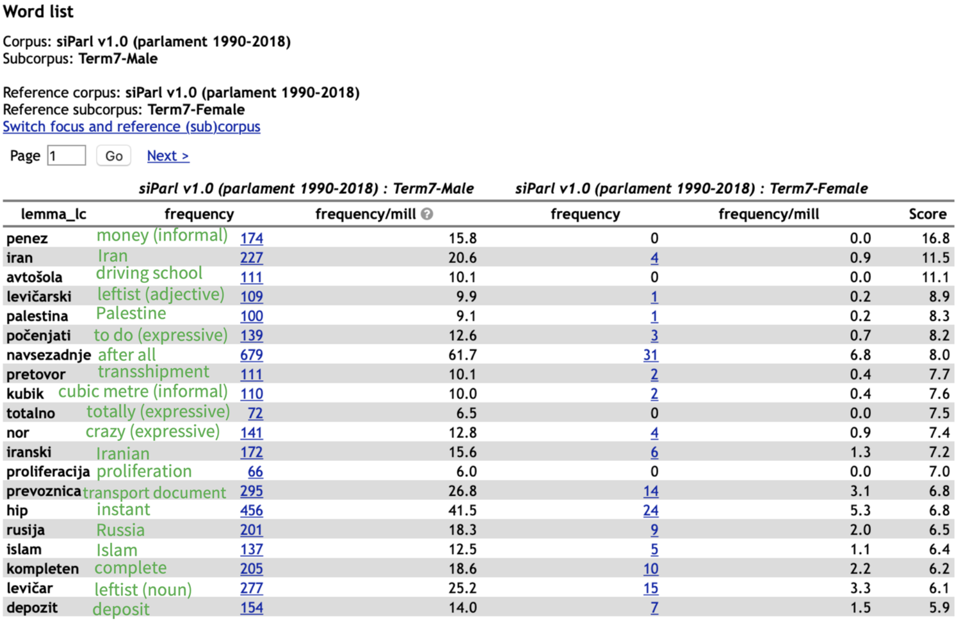
Corpus: **siParl v1.0 (parlament 1990-2018)**
Subcorpus: **Term7-Female**

Reference corpus: **siParl v1.0 (parlament 1990-2018)**
Reference subcorpus: **Term7-Male**
Switch focus and reference (sub)corpus

Page 1 Go Next >

		siParl v1.0 (parlament 1990-2018) : Term7-Female		*siParl v1.0 (parlament 1990-2018) : Term7-Male*		
lemma_lc		frequency	frequency/mill	frequency	frequency/mill	Score
plazma	plasma	79	17.3	4	0.4	13.4
mark	Mark Medical (a company)	124	27.2	16	1.5	11.5
kb	KB (a company)	74	16.2	7	0.6	10.5
medical	Mark Medical (a company)	98	21.5	15	1.4	9.5
psihiatričen	psychiatric	44	9.7	2	0.2	9.0
pediatrija	paediatrics	66	14.5	9	0.8	8.5
rejništvo	foster care	163	35.8	37	3.4	8.4
bunc	Bunc (surname)	32	7.0	0	0.0	8.0
enaročanje	e-appointment service	66	14.5	11	1.0	7.7
pacientov	the patient's	199	43.6	53	4.8	7.7
e-napotnica	e-referral	45	9.9	5	0.5	7.5
vega	Vega Finanz (a company)	36	7.9	3	0.3	7.0
žilen	vascular	392	86.0	126	11.5	7.0
duševen	mental	241	52.9	77	7.0	6.7
posredovalec	responder	59	12.9	12	1.1	6.7
opornica	stenting	397	87.1	139	12.6	6.5
medicala	Mark Medical (genitive case)	24	5.3	0	0.0	6.3
ukc	University Medical Center	667	146.3	251	22.8	6.2
skrajševanje	reducing	144	31.6	48	4.4	6.1
nmp	emergency medical assistance	38	8.3	6	0.5	6.0

Figure 7: Twenty top-ranking keywords of female speakers (with their English translations in green) in the seventh parliamentary term compared to their male counterparts.



Word list

Corpus: **siParl v1.0 (parlament 1990-2018)**
Subcorpus: **Term7-Male**

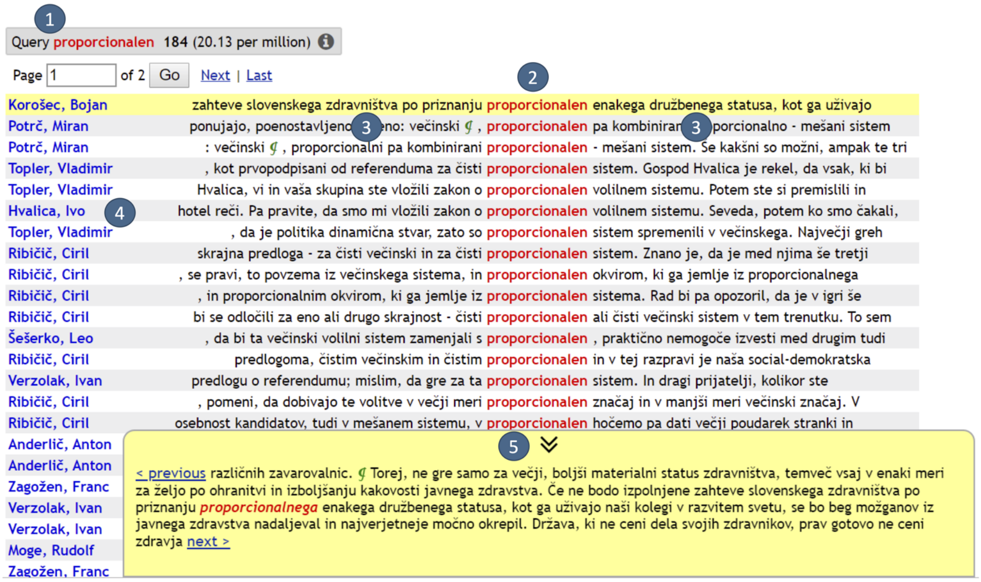
Reference corpus: **siParl v1.0 (parlament 1990-2018)**
Reference subcorpus: **Term7-Female**
Switch focus and reference (sub)corpus

Page 1 Go Next >

		siParl v1.0 (parlament 1990-2018) : Term7-Male		*siParl v1.0 (parlament 1990-2018) : Term7-Female*		
lemma_lc		frequency	frequency/mill	frequency	frequency/mill	Score
penez	money (informal)	174	15.8	0	0.0	16.8
iran	Iran	227	20.6	4	0.9	11.5
avtošola	driving school	111	10.1	0	0.0	11.1
levičarski	leftist (adjective)	109	9.9	1	0.2	8.9
palestina	Palestine	100	9.1	1	0.2	8.3
počenjati	to do (expressive)	139	12.6	3	0.7	8.2
navsezadnje	after all	679	61.7	31	6.8	8.0
pretovor	transshipment	111	10.1	2	0.4	7.7
kubik	cubic metre (informal)	110	10.0	2	0.4	7.6
totalno	totally (expressive)	72	6.5	0	0.0	7.5
nor	crazy (expressive)	141	12.8	4	0.9	7.4
iranski	Iranian	172	15.6	6	1.3	7.2
proliferacija	proliferation	66	6.0	0	0.0	7.0
prevoznica	transport document	295	26.8	14	3.1	6.8
hip	instant	456	41.5	24	5.3	6.8
rusija	Russia	201	18.3	9	2.0	6.5
islam	Islam	137	12.5	5	1.1	6.4
kompleten	complete	205	18.6	10	2.2	6.2
levičar	leftist (noun)	277	25.2	15	3.3	6.1
depozit	deposit	154	14.0	7	1.5	5.9

Figure 8: Twenty top-ranking keywords of male speakers (with their English translations in green) in the seventh parliamentary term compared to their female counterparts.

- Concordances can be displayed directly by clicking on the key lemma in the keyword list. At the top of the screen the query is displayed along with its hits (see *Item 1*).



Query proporcionalen 184 (20.13 per million)

Page 1 of 2 Go Next | Last

Korošec, Bojan	zahteve slovenskega zdravništva po priznanju	proporcionalen	enakega družbenega statusa, kot ga uživajo
Potrč, Miran	ponujajo, poenostavljeno [illegible]eno: večinski ¶ ,	proporcionalen	pa kombiniran[illegible]porcionalno - mešani sistem
Potrč, Miran	: večinski ¶ , proporcionalni pa kombinirani	proporcionalen	- mešani sistem. Še kakšni so možni, ampak te tri
Topler, Vladimir	, kot prvopodpisani od referenduma za čisti	proporcionalen	sistem. Gospod Hvalica je rekel, da vsak, ki bi
Topler, Vladimir	Hvalica, vi in vaša skupina ste vložili zakon o	proporcionalen	volilnem sistemu. Potem ste si premislili in
Hvalica, Ivo	hotel reči. Pa pravite, da smo mi vložili zakon o	proporcionalen	volilnem sistemu. Seveda, potem ko smo čakali,
Topler, Vladimir	, da je politika dinamična stvar, zato so	proporcionalen	sistem spremenili v večinskega. Največji greh
Ribičič, Ciril	skrajna predloga - za čisti večinski in za čisti	proporcionalen	sistem. Znano je, da je med njima še tretji
Ribičič, Ciril	, se pravi, to povzema iz večinskega sistema, in	proporcionalen	okvirom, ki ga jemlje iz proporcionalnega
Ribičič, Ciril	, in proporcionalnim okvirom, ki ga jemlje iz	proporcionalen	sistema. Rad bi pa opozoril, da je v igri še
Ribičič, Ciril	bi se odločili za eno ali drugo skrajnost - čisti	proporcionalen	ali čisti večinski sistem v tem trenutku. To sem
Šešerko, Leo	, da bi ta večinski volilni sistem zamenjali s	proporcionalen	, praktično nemogoče izvesti med drugim tudi
Ribičič, Ciril	predlogoma, čistim večinskim in čistim	proporcionalen	in v tej razpravi je naša social-demokratska
Verzolak, Ivan	predlogu o referendumu; mislim, da gre za ta	proporcionalen	sistem. In dragi prijatelji, kolikor ste
Ribičič, Ciril	, pomeni, da dobivajo te volitve v večji meri	proporcionalen	značaj in v manjši meri večinski značaj. V
Ribičič, Ciril	osebnost kandidatov, tudi v mešanem sistemu, v	proporcionalen	hočemo pa dati večji poudarek stranki in
Anderlič, Anton			
Anderlič, Anton			
Zagožen, Franc			
Verzolak, Ivan			
Verzolak, Ivan			
Moge, Rudolf			
Zagožen, Franc			

< previous različnih zavarovalnic. ¶ Torej, ne gre samo za večji, boljši materialni status zdravništva, temveč vsaj v enaki meri za željo po ohranitvi in izboljšanju kakovosti javnega zdravstva. Če ne bodo izpolnjene zahteve slovenskega zdravništva po priznanju *proporcionalnega* enakega družbenega statusa, kot ga uživajo naši kolegi v razvitem svetu, se bo beg možganov iz javnega zdravstva nadaljeval in najverjetneje močno okrepil. Država, ki ne ceni dela svojih zdravnikov, prav gotovo ne ceni zdravja next >

Figure 9: Extended context for the first hit in the concordance list of the keyword 'proporcionalen/proportional'.

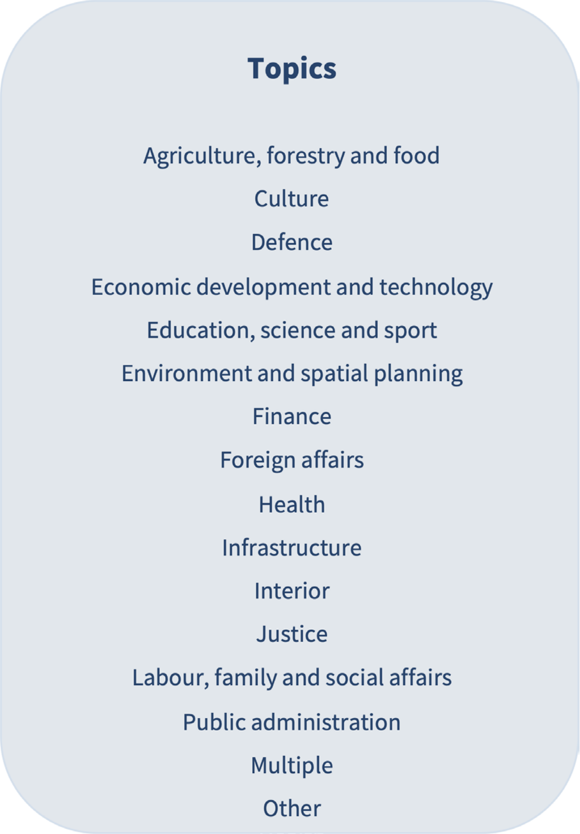
- The words in red at the centre of the screen (see *Item 2*) are the hits of the search word in our subcorpus and the text in black (see *Item 3*) is the context.
- The text in blue on the left (see *Item 4*) is the metadata for the concordances, in our case the speaker.
- The context can be further extended by clicking on the desired concordance (see *Item 5*). The same procedure can be followed for obtaining more metadata by clicking on the speaker.

6.3.3 Comparative analysis

Because the main role of the parliament is legislative, and because the legislative and budgetary discussions are structured according to the ministries responsible for them, we chose to use the list of fourteen ministries of the current Slovenian government as categories for topic analysis. While any other list of topics could be used, this one felt the most natural in the specific setting of parliamentary discourse. The categories are listed in **Table 2**. Illustrative examples of manual annotation of the ten top-ranking keywords by female and by male speakers are given in **Table 3**. Keywords that are used in multiple topics according to concordance analysis were assigned the label *Multiple*. Keywords that could not be assigned a topic because they are interactive or procedural expressions or stylistic devices were assigned the label *Other* and excluded from the rest of the analysis.

Tables 4 and **5** contain summarised results of the manual annotation of 100 top-ranking key lemmas for female and for male speakers in the first and the last parliamentary term. The results show that the range of topics is comparable through time and between the genders. Despite the similar number of identified topics, men and women differ a great deal in their most prominent topics.

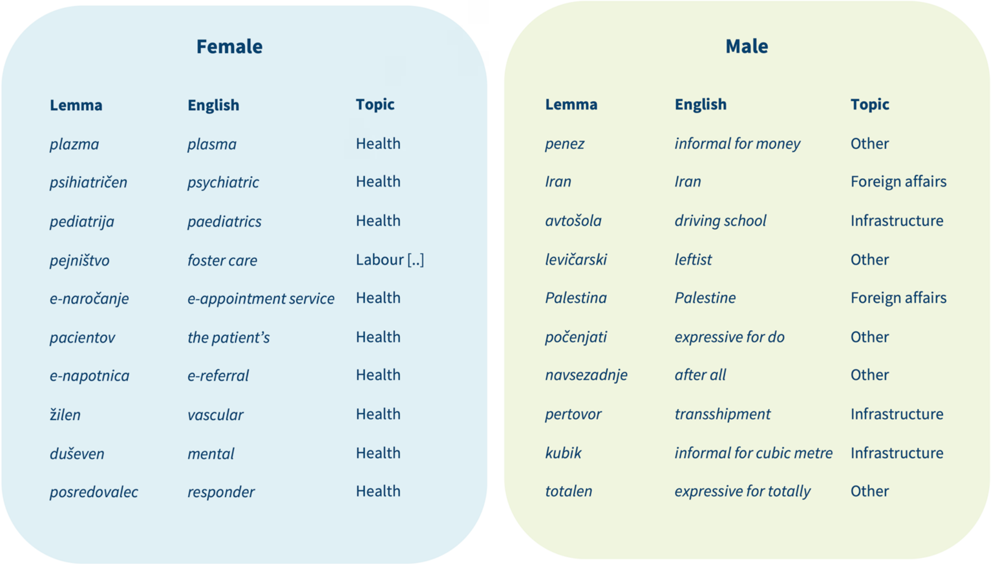
In the first parliamentary term, the majority (60%) of all the analysed female keywords belong to only two topics (*Health* and *Labour, family and social affairs*), whereas the two most prevalent topics of male speakers (*Infrastructure* and *Economic development and technology*) amount to a good third of the sample (37%). In the seventh parliamentary term, the



Topics

Agriculture, forestry and food
Culture
Defence
Economic development and technology
Education, science and sport
Environment and spatial planning
Finance
Foreign affairs
Health
Infrastructure
Interior
Justice
Labour, family and social affairs
Public administration
Multiple
Other

Table 2: Keyword annotation categories.



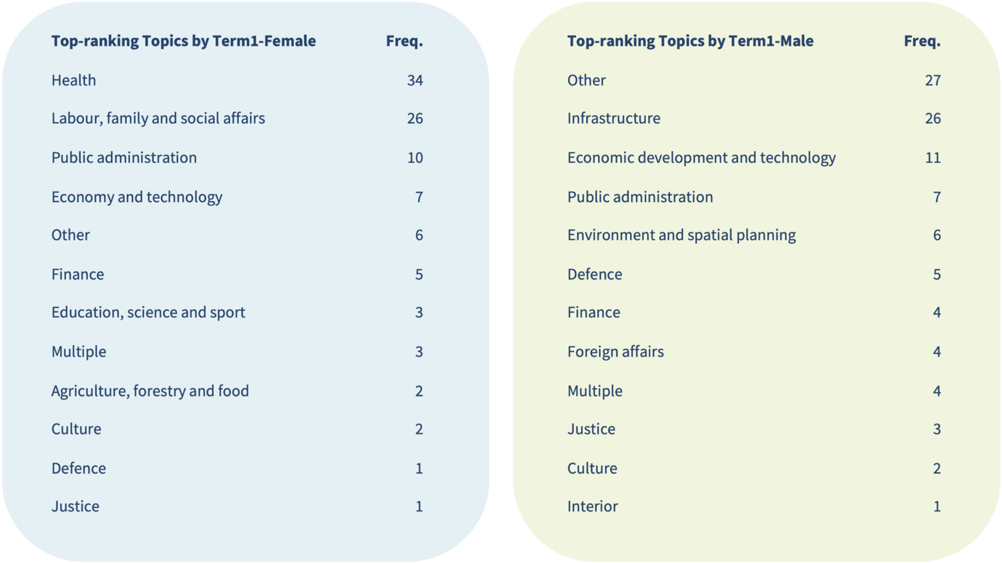
Female

Lemma	English	Topic
plazma	*plasma*	Health
psihiatričen	*psychiatric*	Health
pediatrija	*paediatrics*	Health
pejništvo	*foster care*	Labour [..]
e-naročanje	*e-appointment service*	Health
pacientov	*the patient's*	Health
e-napotnica	*e-referral*	Health
žilen	*vascular*	Health
duševen	*mental*	Health
posredovalec	*responder*	Health

Male

Lemma	English	Topic
penez	*informal for money*	Other
Iran	*Iran*	Foreign affairs
avtošola	*driving school*	Infrastructure
levičarski	*leftist*	Other
Palestina	*Palestine*	Foreign affairs
počenjati	*expressive for do*	Other
navsezadnje	*after all*	Other
pertovor	*transshipment*	Infrastructure
kubik	*informal for cubic metre*	Infrastructure
totalen	*expressive for totally*	Other

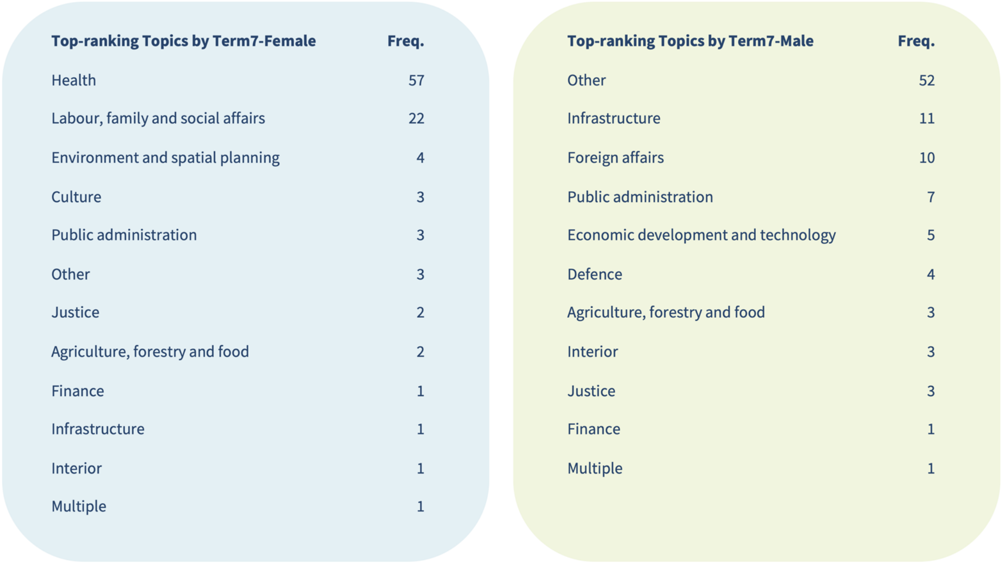
Table 3: Illustrative example of manual topic annotation of ten top-ranking keywords from speeches by female and male speakers in the seventh parliamentary term.



Top-ranking Topics by Term1-Female	Freq.
Health	34
Labour, family and social affairs	26
Public administration	10
Economy and technology	7
Other	6
Finance	5
Education, science and sport	3
Multiple	3
Agriculture, forestry and food	2
Culture	2
Defence	1
Justice	1

Top-ranking Topics by Term1-Male	Freq.
Other	27
Infrastructure	26
Economic development and technology	11
Public administration	7
Environment and spatial planning	6
Defence	5
Finance	4
Foreign affairs	4
Multiple	4
Justice	3
Culture	2
Interior	1

Table 4: Topics of the 100 top-ranking keywords in speeches by female and male speakers in the first parliamentary term.



Top-ranking Topics by Term7-Female	Freq.
Health	57
Labour, family and social affairs	22
Environment and spatial planning	4
Culture	3
Public administration	3
Other	3
Justice	2
Agriculture, forestry and food	2
Finance	1
Infrastructure	1
Interior	1
Multiple	1

Top-ranking Topics by Term7-Male	Freq.
Other	52
Infrastructure	11
Foreign affairs	10
Public administration	7
Economic development and technology	5
Defence	4
Agriculture, forestry and food	3
Interior	3
Justice	3
Finance	1
Multiple	1

Table 5: Topics of the 100 top-ranking keywords in speeches by female and male speakers in the seventh parliamentary term.

two prevailing female topics not only remained the same but also further intensified (79%), while in men we recorded a shift from *Economic development and technology* towards *Foreign affairs*. While the division of topics between genders cannot be explained through corpus analysis, both an intensified focus on *Health* and *Labour, family and social affairs* in the recent term and the shift from *Economy* to *Foreign affairs* are a good reflection of the state of things at the time of independence, when the entire Yugoslavian market was lost and the economy had to shift from a socialist to capitalist regime and therefore required intensive discussions in the first parliamentary term. The last parliamentary term, on the other hand, is marked by more intensive international trade as well as by greater international security threats which warrant legislative and budgetary decisions in the parliament. The intensified discussions on social issues in the seventh parliamentary term are largely due to heavy pressure on the budget due to the severe economic crisis that Slovenia faced in the period that coincided with that parliamentary term, while health-related issues escalated due to a crumbling public health system.

We can also observe that in both parliamentary terms men and women share only three prominent topics: *Finance*, *Justice* and *Public administration*. Female-specific topics are *Health*, *Labour, family and social affairs* and *Education, science and sport* (only in the first parliamentary term). The only male-specific topic, on the other hand, is *Foreign affairs*. Surprisingly, *Education* disappears from the list of prominent female-specific topics between the first and seventh parliamentary terms. In addition, none of the top-ranking keywords from female speeches in the seventh parliamentary term belongs to the categories of *Defence* or *Economy*. The reverse trend can be observed for the top-ranking male keywords belonging to *Agriculture, forestry and food* which appear in the seventh parliamentary term only.

Our findings point to significant differences in the roles and interests of male and female speakers in the Slovenian parliament, which is in line with previous studies showing that women focus more than men on the so-called soft topics. Diachronic comparisons reveal shifts in both directions: on the one hand, topics of *Health* and *Labour, family and social affairs* have been reinforced as female-specific topics over time and the same can be observed for the male-specific topic of *Foreign affairs*. On the other hand, *Education, science and sport* has disappeared from the list of female-specific topics. Similarly, infrastructural, environmental and safety issues have recently cropped up among top-ranking female keywords, indicating their participation in such discussions as well.

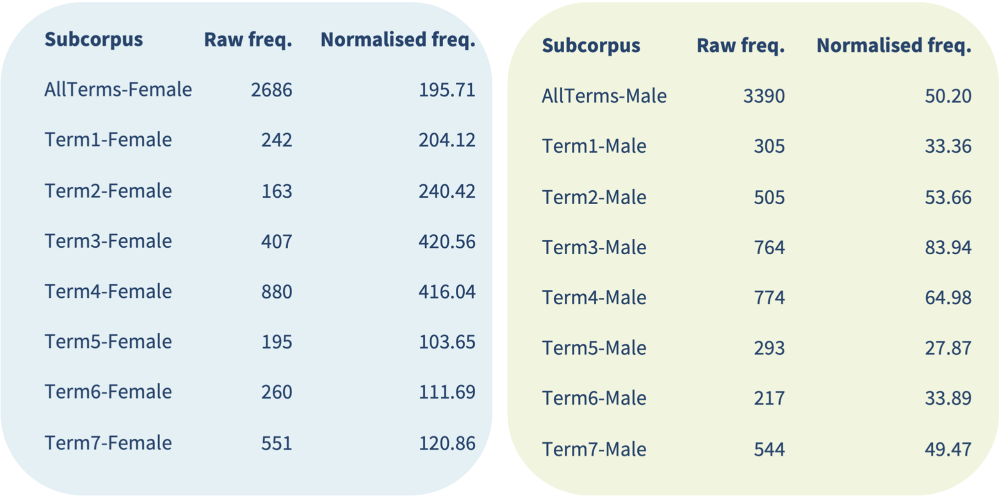
6.4 TASK 3: Women's issues

This task is inspired by related work (see Blaxill & Beelen) which investigated how frequently, by whom and in what way women's issues, such as women's rights, equality, discrimination and so forth, are addressed in parliamentary history. It is interesting that the impact of gender seems to be prominent even in countries with high representation of women in the parliament. Bäck et al., for example, found that female MPs in the Swedish parliament discuss so-called hard policy issues less often. Furthermore, Antić Gaber states that society usually expects female MPs to be actively involved in different policy areas than men. Those issues, often referred to as women's issues in political science, are defined as pertaining to policy areas that are particularly salient to women because of women's historical role in society or because those areas directly affect women's lives.

In Task 3 we are interested in comparing how male and female speakers in the Slovenian parliament express themselves when addressing women's issues by focusing on their use of the noun 'ženska/female' as an explicit indicator of debate on women's issues.

6.4.1 Working with frequencies

First, we are interested in how frequently the noun 'ženska/female' is mentioned by male and female speakers in different parliamentary terms from 1992 to 2018. We query all the subcorpora for the lemma and record all the frequency counts in a spreadsheet, as can be seen in **Table 6**. Because we are interested in comparing subcorpora of different sizes, it is important to use normalised frequencies instead of the raw ones, as raw frequencies can be misleading. For instance, if we only looked at raw frequencies of the lemmas by female (407) and by



Subcorpus	Raw freq.	Normalised freq.
AllTerms-Female	2686	195.71
Term1-Female	242	204.12
Term2-Female	163	240.42
Term3-Female	407	420.56
Term4-Female	880	416.04
Term5-Female	195	103.65
Term6-Female	260	111.69
Term7-Female	551	120.86

Subcorpus	Raw freq.	Normalised freq.
AllTerms-Male	3390	50.20
Term1-Male	305	33.36
Term2-Male	505	53.66
Term3-Male	764	83.94
Term4-Male	774	64.98
Term5-Male	293	27.87
Term6-Male	217	33.89
Term7-Male	544	49.47

Table 6: Absolute and relative frequencies of mentions of the lemma 'ženska/female' in siParl.

male speakers (764) in the third parliamentary term, we could conclude that male speakers mention the noun 'ženska/female' nearly twice as often than their female counterparts. But, because the female subcorpus in that parliamentary term is much smaller than the male subcorpus, the normalised frequency, which calculates frequency per the same number of words in each subcorpus, shows that the noun in question is actually five times more frequent in the female subcorpus (420.56) compared to the male one (83.94).

Overall, normalised frequency of the word for the entire period is nearly four times higher in the subcorpus of female speakers compared to their male counterparts (195.71 vs. 50.2). As can be seen from **Figure 10**, female speakers talked about women considerably more than their male counterparts until about 2010, but this difference shrunk considerably and stabilised after the fifth parliamentary term. The most striking results are observed in the third and fifth parliamentary terms. In the third parliamentary term, the normalised frequency of the word nearly doubled in both subcorpora, which might be a consequence of record low numbers of female representatives in previous parliamentary terms and their unequal position in the society overall. A decade later, in the fifth parliamentary term there was a sudden drop in the number of mentions of the word in both subcorpora, which is especially pronounced in the female corpus where it fell to a quarter of its previous frequency. This coincides with the period of a major global economic crisis which badly hit Slovenia and probably took centre-stage in parliamentary discussions, but this would need to be confirmed through further investigation and contextualisation using qualitative methods, such as concordance analysis.

6.4.2 Extracting collocations

Next, we will demonstrate another popular corpus analysis technique called collocation extraction which identifies word combinations that co-occur more often than would be expected by chance. While collocations are most typically employed in lexicography and related fields of applied linguistics, we will use them as a vehicle to explore the concepts or themes that are debated in the parliament.

To be able to dive deeper into the issues addressed when talking about women, we will investigate collocations of the noun 'ženska/female' from two siParl subcorpora, one containing

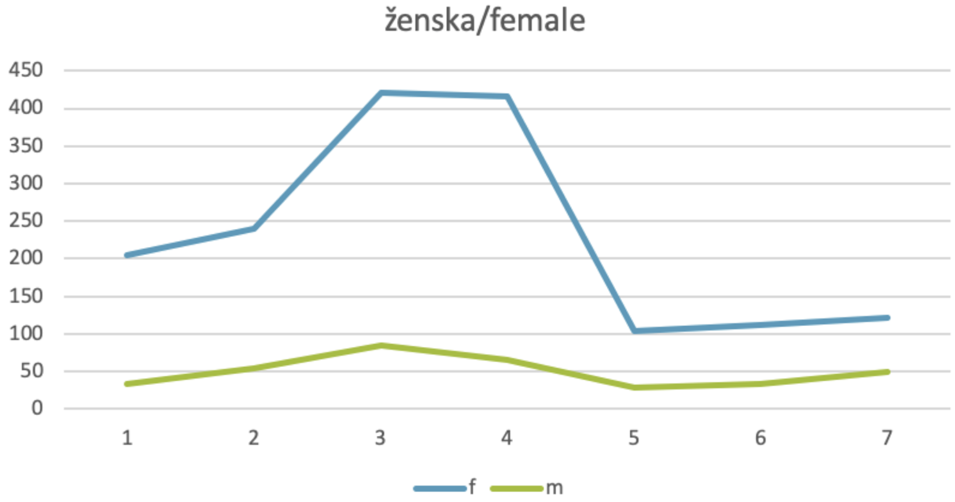


Figure 10: Normalised frequency of mentions of 'ženska/female' by female and male speakers in siParl.

the female speeches from all seven parliamentary terms (*AllTerms-Female*) and the other one comprising male speeches from the same timespan (*AllTerms-Male*).

A screencast of how to extract collocations in noSketch Engine is available here.

We extracted collocations in the range one word to the left and one word to the right of the headword, with five occurrences as the minimum frequency in the corpus and three minimum co-occurrences with the headword in the defined window size. While window size and minimum frequencies can be set manually and depend on type and frequency of the word under investigation, corpus size and our research goal, we wanted to limit our analysis for this tutorial to fixed multi-word expressions, thereby using a very narrow window and strict minimum frequency criteria.

We used the logDice statistic measure to measure association strength between words. While several other collocation measures are also offered by the noSketch Engine concordancer, such as Mutual Information or T-score, we opted for logDice because it is not affected by the size of the corpus and can therefore be used to compare the scores between subcorpora of different sizes.

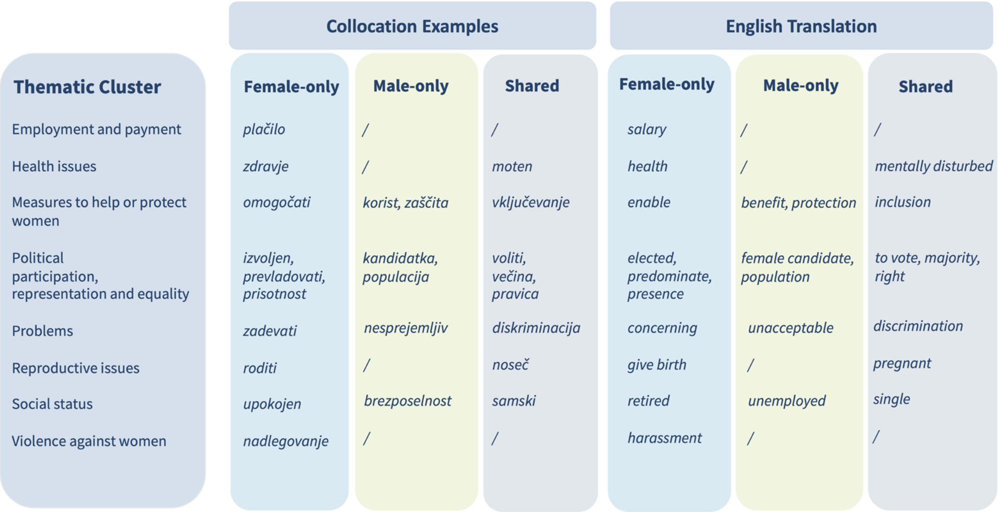
The two collocation lists for the headword 'ženska/female' are available here:

- AllTerms-Female
- AllTerms-Male

Both collocation lists were imported into a spreadsheet and manually analysed.

6.4.3 Comparative analysis

We took 100 top-ranking collocation candidates from each list and manually divided them into three categories: female only, male only and shared. Next, we categorised each collocation candidate into one of eight thematic clusters that were inspired from close reading of corpus concordances as illustrated in **Table 7**.



	Collocation Examples			English Translation		
Thematic Cluster	Female-only	Male-only	Shared	Female-only	Male-only	Shared
Employment and payment	plačilo	/	/	salary	/	/
Health issues	zdravje	/	moten	health	/	mentally disturbed
Measures to help or protect women	omogočati	korist, zaščita	vključevanje	enable	benefit, protection	inclusion
Political participation, representation and equality	izvoljen, prevladovati, prisotnost	kandidatka, populacija	voliti, večina, pravica	elected, predominate, presence	female candidate, population	to vote, majority, right
Problems	zadevati	nesprejemljiv	diskriminacija	concerning	unacceptable	discrimination
Reproductive issues	roditi	/	noseč	give birth	/	pregnant
Social status	upokojen	brezposelnost	samski	retired	unemployed	single
Violence against women	nadlegovanje	/	/	harassment	/	/

Table 7: Examples of collocation clusters.

The results show that more than half (55%) of the collocations are shared between speakers of both genders, revealing common ground in the understanding of the women's position in modern society. The shared collocations mainly refer to concepts related to social status (e.g. 'samski/single'), representation (e.g. 'participacija/participation'), equality (e.g. 'emancipacija/emancipation') and reproduction (e.g. 'neploden/infertile').

Judging from their frequent use of collocations with a negative connotation, male speakers focused on the unsolved problems in the society (e.g. 'zatiranje/oppression') and tend to place women in a more passive position, especially when the context included measures to help or protect women (e.g. 'vključiti/to include', 'spraviti/to get'). The male-only collocations are also often related to outstanding problems (e.g. 'zadaj/behind') or women's social status and elections (e.g. 'poročen/married', 'kandidatka/female candidate'). Female speakers, on the other hand, often used words referencing their self-agency or independence (e.g. 'ambiciozen/ambitious'), but also problems such as their inequality (e.g. 'vključenost/inclusion'), social status (e.g. 'upokojen/retired') and aggression (e.g. 'nadlegovanje/harassment'). Our results are in line with the related work.

These results confirm the findings of Antić Gaber, who looked at a shorter period (eight years) of female MPs' activity in the Slovene parliament and showed with diverse, non-corpus linguistic methods that there exists a clear difference between legislative priorities of male and female MPs, identifying the same topics as our analysis.

7 Conclusions

The aim of this tutorial was to demonstrate the potential of linguistic corpora and corpus analysis techniques for the analysis of socio-cultural phenomena and trends observed through language use in specialised discourse. We have shown how methods of corpus linguistics enable quantitative as well as qualitative observations that go beyond the researcher's intuition and thus offer greater transparency, objectivity, reliability and replicability, which are becoming increasingly important in data-driven humanities and social science research.

The contribution of this tutorial is three-fold. First, we have demonstrated the importance of understanding the content and structure of a research dataset in order to be able to maximise its potential for our research. Second, we have showcased how a set of standard corpus analysis techniques can be utilised well beyond quantification only and simple corpus queries. Instead, we have systematically used the output provided by the concordancer as a starting point for a detailed qualitative manual analysis that carefully situates the results in the relevant socio-linguistic context. Last but not least, we have situated the tutorial in a real-life research setting, demonstrating the application of common corpus analysis techniques to tackle a set of trending research questions in humanities and social science.

While the tutorial is based on the corpus of Slovenian parliamentary debates, students and scholars are strongly encouraged to replicate the analyses using parliamentary corpora for other languages, thereby contributing to the multilingual, transnational and transcultural comparative research of parliamentary discourse.

Acknowledgements

The work described in this paper was funded by the Horizon 2020 project PARTHENOS – Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies (Grant Agreement n. 654119), and by the Horizon 2020 project SSHOC – Social Sciences and Humanities Open Cloud (Grant Agreement n. 823782).

Author Information

Darja Fišer is Associate Professor at the Department of Translation Studies at the Faculty of Arts, University of Ljubljana, and Senior Research Fellow at the Department of Knowledge Technologies at the Jožef Stefan Institute. She is also Vice Executive Director of CLARIN ERIC, a European Research Infrastructure for Language Resources and Language Technologies, and Chair of the Steering Committee of ESSLLI, the oldest and biggest European summer school on language, logic and information. As a researcher, she is currently active in the fields of computer-mediated communication and socially unacceptable online discourse practices using corpus-linguistics methods and natural language processing.

Kristina Pahor de Maiti is a Research Assistant at the Department of Translation Studies of the Faculty of Arts, University of Ljubljana, and a PhD student at the Department of Comparative and General Linguistics at the same faculty. She earned her MA in Interpretation studies at the University of Ljubljana and later on worked as a translator and interpreter in the private sector. Her research interests include spoken language and computer-mediated communication. Currently, she is focusing on corpus-based analysis of socially unacceptable discourse online.

Videos

Videos from all the tutorials in this collection have been archived in a special playlist on the Digital Modern Languages YouTube channel at: https://www.youtube.com/playlist?list=PLfaB2f0CyBdu1OluU36KKLKXpKWDco62q

Here is a list of videos associated with this tutorial:

- Creating subcorpora in noSketch Engine https://www.youtube.com/watch?v=PanYwJw-_Mo
- Creating frequency lists in noSketch Engine https://www.youtube.com/watch?v=oZh-GoNqfIM
- Extracting keywords in noSketch Engine https://www.youtube.com/watch?v=x1QLeB0Z8P0
- Extracting collocations in noSketch Engine https://www.youtube.com/watch?v=F-dzi47_p5w

References

Alasuutari, Pertti, et al. 'The Rise of the Idea of Model in Policymaking: The Case of the British Parliament', 1803–2005 *European Journal of Sociology/Archives Européennes de Sociologie*, vol. 59, no. 3, 2018, pp. 341–63. DOI: https://doi.org/10.1017/S0003975618000164

Antić Gaber, Milica. 'Women in the Slovene Parliament: Working towards Critical Mass', *Women in East European Politics*, 2004, pp. 19–32.

Bäck, Hanna, et al. *Who Takes the Parliamentary Floor? The Role of Gender in Speech-Making in the Swedish Riksdag.* 2014. Accessed 4 May 2020. DOI: https://doi.org/10.1177/1065912914525861

Baker, Paul. *Using Corpora to Analyze Gender*. A&C Black, 2014.

Biber, Douglas, and Randi Reppen. *The Cambridge Handbook of English Corpus Linguistics*. Cambridge University Press, 2015. DOI: https://doi.org/10.1017/CBO9781139764377

Biel, Łucja, et al. 'Collocations of Terms in EU Competition Law: A Corpus Analysis of EU English Collocations', *Language and Law*, Springer, 2018, pp. 249–74. DOI: https://doi.org/10.1007/978-3-319-90905-9_14

Blaxill, L., and K. Beelen. *Women in Parliament Since 1945: Have They Changed the Debate?* History & Policy, 2016.

Coates, Jennifer. *Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language*. Longman, 1997.

Demmen, Jane, et al. 'Charting the Semantics of Labour Relations in House of Commons Debates Spanning Two Hundred Years', *Doing Politics: Discursivity, Performativity and Mediation in Political Discourse*, vol. 80, 2018, p. 81. DOI: https://doi.org/10.1075/dapsac.80.04dem

Eckert, Penelope, and Sally McConnell-Ginet. *Language and Gender*. Cambridge University Press, 2013. DOI: https://doi.org/10.1017/CBO9781139245883

Goddard, Angela, and Lindsey Meân Patterson. *Language and Gender*. London, New York: Routledge, 2015.

Hansen, Dorte Haltrup, et al. 'A Pilot Gender Study of the Danish Parliament Corpus', *Proceedings of the ParlaClarin Workshop at the Eleventh International Conference on Language Resources and Evaluation: (LREC 2018)*. European Language Resources Association. 2018.

Ilie, Cornelia. 'Parliamentary Discourse', *The International Encyclopedia of Language and Social Interaction*, 2015, pp. 1–15. DOI: https://doi.org/10.1002/9781118611463.wbielsi201

Jaworska, Sylvia, and Kath Ryan. 'Gender and the Language of Pain in Chronic and Terminal Illness: A Corpus-Based Discourse Analysis of Patients' Narratives', *Social Science & Medicine*, vol. 215, 2018, pp. 107–14. DOI: https://doi.org/10.1016/j.socscimed.2018.09.002

Karan, Mladen, et al. 'Analysis of Policy Agendas: Lessons Learned from Automatic Topic Classification of Croatian Political Texts', *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.* Association for Computational Linguistics. 2016, pp. 12–21. DOI: https://doi.org/10.18653/v1/W16-2102

Leijenaar, Monique. *How to Create a Gender Balance in Political Decision-Making: A Guide to Implementing Policies for Increasing the Participation of Women in Political Decision-Making*. Office for Official Publications of the European Communities, 1997.

Litosseliti, L. *Gender and Language: Theory and Practice*. Hodder Arnold, 2006.

Marshall, Catherine. 'Policy Discourse Analysis: Negotiating Gender Equity', *Journal of Education Policy*, vol. 15, no. 2, 2000, pp. 125–56. DOI: https://doi.org/10.1080/026809300285863

McEnery, Tony, and Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, 2011. DOI: https://doi.org/10.1017/CBO9780511981395

Mensah, Eric Opoku, and Sandra Freda Wood. 'Articulations of Feminine Voices in Ghana's Parliament: A Study of the Hansard from 2010–2011', *AFRREV LALIGENS: An International Journal of Language, Literature and Gender Studies*, vol. 7, no. 2, 2018, pp. 61–77. DOI: https://doi.org/10.4314/laligens.v7i2.6

Mollin, Sandra. 'The Hansard Hazard: Gauging the Accuracy of British Parliamentary Transcripts', *Corpora*, vol. 2, no. 2, 2007, pp. 187–210. DOI: https://doi.org/10.3366/cor.2007.2.2.187

Newman, Matthew L., et al. 'Gender Differences in Language Use: An Analysis of 14,000 Text Samples', *Discourse Processes*, vol. 45, no. 3, 2008, pp. 211–36. DOI: https://doi.org/10.1080/01638530802073712

Osborn, Tracy L. *How Women Represent Women: Political Parties, Gender and Representation in the State Legislatures*. Oxford University Press, 2012. DOI: https://doi.org/10.1093/acprof:oso/9780199845347.001.0001

Pančur, Andrej, et al. *Slovenian Parliamentary Corpus SiParl 1.0 (1990–2018)*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1236**.** 2019.

Rix, Kathryn. '"Whatever Passed in Parliament Ought to Be Communicated to the Public": Reporting the Proceedings of the Reformed Commons, 1833–50', *Parliamentary History*, vol. 33, no. 3, 2014, pp. 453–74. DOI: https://doi.org/10.1111/1750-0206.12106

Selišnik, Irena, and Milica Antić Gaber. 'From Voluntary Party to Legal Electoral Gender Quotas in Slovenia: The Importance and Limitations of Legal and Institutional Mechanisms', *EUI Department of Law Research Paper*, no. 2015/31, 2015. DOI: https://doi.org/10.2139/ssrn.2610662

Wängnerud, Lena. 'Intressen Kontra Stereotyper. Varför Finns Det Kvinnliga Och Manliga Politikområden i Riksdagen?' *Statsvetenskaplig Tidskrift*, vol. 99, no. 2, 1996.

Wodak, Ruth. 'Pragmatics and Critical Discourse Analysis: A Cross-Disciplinary Inquiry', *Pragmatics & Cognition*, vol. 15, Apr. 2007, pp. 203–25. DOI: https://doi.org/10.1075/pc.15.1.13wod

Wolbrecht, Christina. *Female Legislators and the Women's Rights Agenda: From Feminine Mystique to Feminist Era*. University of Oklahoma Press, Norman, 2002, pp. 170–97.

How to cite this article: Fišer, D and Pahor de Maiti, K 2020 Voices of the Parliament. *Modern Languages Open,* 2020(1): 46 pp. 1–23. DOI: https://doi.org/10.3828/mlo.v0i0.295

Published: 28 August 2020

]u[*Modern Languages Open* is a peer-reviewed open access journal published by Liverpool University Press.

OPEN ACCESS