**VOICES OF THE PARLIAMENT**

**Abstract**

This tutorial shows how corpora can be used to investigate language use and communication practices in a specialized socio-cultural context of political discourse in order to explore socio-cultural phenomena. 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 language use in the Slovenian Parliament.

**Table of contents**

[1 Introduction 2](#_Toc57790732)

[2 Instructions for use 2](#_Toc57790733)

[3 Corpora and concordancers 3](#_Toc57790734)

[3.1 Corpora 3](#_Toc57790735)

[3.2 Concordancers 4](#_Toc57790736)

[4 Parliamentary records 5](#_Toc57790737)

[4.1 Parliamentary discourse 5](#_Toc57790738)

[4.2 Faithfulness of the records 5](#_Toc57790739)

[4.3 Know your research dataset 6](#_Toc57790740)

[5 Language and gender 6](#_Toc57790741)

[6 Corpus analysis 7](#_Toc57790742)

[6.1 The siParl 2.0 corpus 7](#_Toc57790743)

[6.2 TASK 1: Representation of women in the Slovenian parliament 9](#_Toc57790744)

[6.2.1 Creating subcorpora 9](#_Toc57790745)

[6.2.2 Using frequency lists 10](#_Toc57790746)

[6.2.3 Comparative analysis 12](#_Toc57790747)

[6.3 TASK 2: Issues addressed by women 13](#_Toc57790748)

[6.3.1 Extracting keywords 14](#_Toc57790749)

[6.3.2 Analyzing concordances 15](#_Toc57790750)

[6.3.3 Comparative analysis 16](#_Toc57790751)

[6.4 TASK 3: Topics related to women 19](#_Toc57790752)

[6.4.1 Working with frequencies 20](#_Toc57790753)

[6.4.2 Extracting collocations 22](#_Toc57790754)

[6.4.3 Comparative analysis 23](#_Toc57790755)

[7 Conclusions 25](#_Toc57790756)

[8 Acknowledgements 25](#_Toc57790757)

[9 References 25](#_Toc57790758)

# Introduction

Parliaments represent the main fora for political debate that shapes legislation which directly impacts our everyday lives. 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 citizens (Ilie, 2017). This is also why parliaments have always been of interest to scholars from a range of disciplines in the humanities and social science.

Parliamentary proceedings are being increasingly made available in a digitized form, and 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 these to perform diverse linguistic, stylistic, cultural, societal and political studies (Biel et al., 2018; Jaworska and Ryan, 2018).

While corpus methods are widely used in linguistics (Biber and Reppen, 2015; McEnery and Hardie, 2011), including gender analysis (Baker, 2014), 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, as well as 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 2.0 corpus](http://hdl.handle.net/11356/1300) which contains parliamentary debates of the National Assembly of the Republic of Slovenia from 1990 to 2018, although knowledge of Slovenian is not 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](https://www.clarin.eu/resource-families/parliamentary-corpora#introduction).

# Instructions for use

This tutorial starts with a brief introduction to corpora and corpus analysis, followed by an introduction of the characteristics of specialized corpora of parliamentary debates and an overview of research into language and gender. The second part of the tutorial is a hands-on, 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 MPs debate in in the Slovenian Parliament over time and to compare and contrast their language use with that 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 [NoSketchEngine](https://nlp.fi.muni.cz/trac/noske) concordancer, while additional manual analysis and visualization of the results will be performed in a spreadsheet editor (e.g., [Google Spreadsheet](https://www.google.com/sheets/about/) or [MS Excel](https://products.office.com/en/excel)).

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

The siParl 2.0 corpus can be queried online through the [NoSketchEngine](https://www.clarin.si/noske/) or [KonText](https://www.clarin.si/kontext/first_form?corpname=siparl20) concordancers at [CLARIN.SI](http://clarin.si/), the Slovenian node of [CLARIN ERIC](https://www.clarin.eu/), the European research infrastructure for language resources and technology. The siParl 2.0 corpus can also be [downloaded](https://www.clarin.si/repository/xmlui/handle/11356/1300) 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](https://programminghistorian.org/en/lessons/corpus-analysis-with-antconc) and [Basic Text Processing in R](https://programminghistorian.org/en/lessons/basic-text-processing-in-r).

This tutorial is an updated version of the [original tutorial which was based on the previous version of the siParl corpus](https://www.modernlanguagesopen.org/articles/10.3828/mlo.v0i0.295/). In comparison to [siParl 1.0](http://hdl.handle.net/11356/1236), the siParl 2.0 corpus contains richer and cleaner speaker and session metadata which makes it possible to distinguish between MPs and other speakers. In addition, speeches have been labelled with parliamentary terms, which simplifies comparative analysis across different legislative periods. Furthermore, additional linguistic annotation layers, such as Universal Dependency features, syntactic parses and named entities, have also been added to the corpus, but since these will not be used in this tutorial, we do not elaborate on them further.

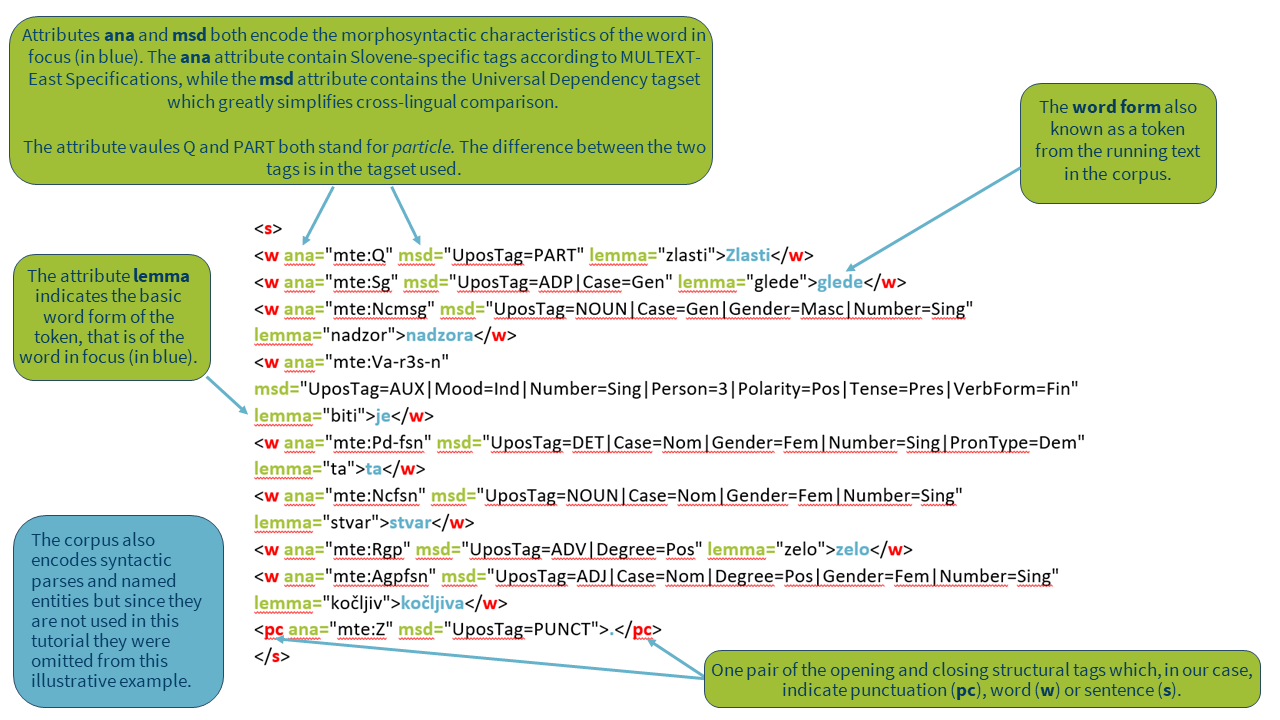
If you wish to immediately proceed to hands-on exercises, skip to section 6 Corpus analysis, since the following three sections (Corpora and concordancers, Parliamentary records and Language and gender) are dedicated to a general theoretical overview and an introduction to the basic terminology. Even though you can follow the tasks in section 6 without studying these three introductory sections, we strongly encourage you to do so before finishing this tutorial. They will give you the necessary theoretical foundations that will ensure a comprehensive understanding and independent use of the demonstrated analytical procedures and adequate interpretation of the results.

# Corpora and concordancers

## Corpora

[Language corpora](https://en.wikipedia.org/wiki/Text_corpus) 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](https://www.w3schools.com/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](https://tei-c.org/).

The basic unit in the corpus is the [token](https://www.sketchengine.eu/my_keywords/token/), the output of an automatic tokenization process which separates a raw text into smaller units, in our case words. The most common way of forming tokens is based on space between the words as a delimiter. Texts in corpora are linguistically annotated to facilitate corpus searches. The two most basic forms of linguistic annotation are [part-of-speech tagging](https://en.wikipedia.org/wiki/Part-of-speech_tagging), which marks up the words in running texts with their part-of-speech, and [lemmatization](https://en.wikipedia.org/wiki/Lemmatisation), which is the assignment of a base form or [lemma](https://en.wikipedia.org/wiki/Lemma_(morphology)) to tokens (word forms). Lemmatization is especially important for corpora of morphologically rich languages, such as Slovenian, while less needed for weakly inflective languages, such as English. In addition to linguistic annotation, corpora are also enriched with metadata about the text (e.g., date, type) and speakers (e.g., name, gender) that are needed for the contextualization of the search results but can also be used for more fine-grained corpus queries.



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

## Concordancers

Corpora are queried with specialized corpus analysis tools, also called [concordancers](http://corpora.lancs.ac.uk/clmtp/2-conc.php). 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](http://martinweisser.org/corpora_site/concordancers.html) detailed list for more examples) but some of the most popular ones are the offline [AntConc](http://www.laurenceanthony.net/software/antconc/) (free) and [WordSmith Tools](http://www.lexically.net/wordsmith/index.html) (for which a license needs to be purchased), where you have to load your own corpora, and the online [BYU concordancer](https://www.english-corpora.org/) (free registration required for full functionality) and [SketchEngine](https://www.sketchengine.eu/documentation/corpus-querying/) (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](https://www.sketchengine.eu/guide/concordance-a-tool-to-search-a-corpus/) show all the hits of the searched word or phrase in the corpus together with its context. They can be randomized or sorted according to the searched word or phrase or by its left or right context, revealing typical patterns in which it is used.
* [Frequency lists](https://www.sketchengine.eu/user-guide/word-list/) summarize the frequencies of all the hits that correspond to the corpus query (either based on a selected word (phrase) or metadata) out of context, and can be sorted alphabetically or by frequency.
* [Keyword lists](https://www.sketchengine.eu/guide/keywords-and-term-extraction/) highlight what words are prominent in a focus corpus compared to a reference one.
* [Collocation lists](https://www.sketchengine.eu/my_keywords/collocation/) return a series of words that co-occur with the searched word more often than would be expected by chance.

Dedicated tutorials for concordancers (e.g., the [SketchEngine Quick Start Guide](https://www.sketchengine.eu/quick-start-guide/)) as well as general corpus-linguistic courses (e.g., the course [Corpus Linguistics: Method, Analysis, Interpretation](https://www.futurelearn.com/courses/corpus-linguistics)) are already available online, and we refer to them wherever possible. The focus of this tutorial, however, is to showcase how the functionalities of a popular concordance tool can be utilized and combined on a specialized corpus of parliamentary data in order to answer real-world research questions in a methodologically sound way, as such skills are still lacking, especially for students and scholars of (digital) humanities and social sciences who are interested in the study of socio-cultural phenomena through language use.

# 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](https://repository.clarin.dk/repository/xmlui/handle/20.500.12115/8)), they are sometimes also accessible as audio/video recordings coupled with corresponding transcriptions (e.g., [Czech Parliamentary Meetings](https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0005-CF9C-4)). When used as a [diachronic](https://en.wikipedia.org/wiki/Synchrony_and_diachrony) source, a parliamentary corpus enables in-depth research of linguistic and societal change over time. Most parliamentary corpora have rich [metadata](https://en.wikipedia.org/wiki/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., 2018; Demmen et al., 2018).

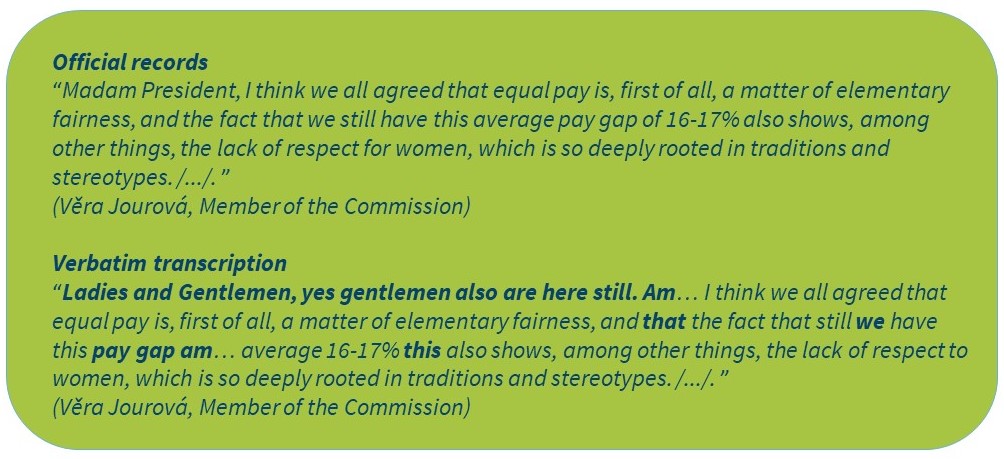
## Parliamentary discourse

Parliaments are institutions governed by specific rules and conventions (see for example [Erskine May](https://erskinemay.parliament.uk/), a description of the UK parliamentary procedure). These are shaped by sociohistorical traditions that influence the organization and operation of the parliament and 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; e.g., interruptions are strictly prohibited in the Greek Parliament, whereas in the UK Parliament these are common and often not sanctioned (Ilie, 2017), which needs to be taken into consideration when forming queries or interpreting the results. Therefore, whenever this type of [discourse](https://en.wikipedia.org/wiki/Discourse) is investigated, it is particularly important to understand the circumstances it was produced in.

## Faithfulness of the records

It should be noted that officially released records of parliamentary debates are not 100% verbatim [transcriptions](https://en.wikipedia.org/wiki/Transcription_(linguistics)), 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, 2007; Rix, 2014). Furthermore, speeches by MPs are often, although not in all parliamentary traditions and all types of parliamentary sessions (interpellations, debates, question time, etc.) prepared in advance (i.e. written-to-be-spoken), which has a big impact on their stylistic features. These peculiarities together with the broader sociohistorical 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](https://en.wikipedia.org/wiki/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 1/5/2017](http://www.europarl.europa.eu/doceo/document/CRE-8-2017-03-01-ITM-019_EN.html) published on the website of the EU Parliament, together with a verbatim transcription of the [video clip](https://www.europarl.europa.eu/plenary/sl/vod.html?mode=unit&vodLanguage=SL&playerStartTime=20170301-20:04:18&playerEndTime=20170301-20:09:42) 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.



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

## Know your research dataset

In order to make the 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](https://exmaralda.org/en/linguistic-annotation-wiki-en/) 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](http://www.chlg.ac.uk/helipad/)), inside the concordancer through which it is made available (e.g., [Hansard in the concordancer of Brigham Young University](https://www.english-corpora.org/hansard/)) or in the repository where the corpus is archived (e.g., [ParlaMeter-hr in the CLARIN.SI repository](https://www.clarin.si/repository/xmlui/handle/11356/1209)). You can read more about the encoding and annotation of parliamentary data in the [Parthenos training module on the collections of parliamentary records](https://training.parthenos-project.eu/sample-page/digital-humanities-research-questions-and-methods/researching-parliamentary-records-in-the-digital-humanities/working-with-parliamentary-corpora/).

# 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 and McConnell-Ginet, 2013; Wodak, 2007). The results show that the differences are subtle but systematic (Newman et al., 2008). Despite this, political communication studies have traditionally been based on male politicians, and only recently have scholars begun to consider the discourse of female politicians (see Marshall, 2000).

This is important because several authors (Antić Gaber and Ilonszki, 2003; Leijenaar, 1997; Wolbrecht, 2002) have demonstrated that female legislators differ from their male counterparts in the issues they address, the positions they take, and the approaches they use in law-making. In their analysis of the representation of women in the UK Parliament after 1945, Blaxill and Beelen (2016) showed a stronger emphasis on the topics related to women in speeches delivered by female MPs, who on average also contributed considerably more speeches about women in comparison to male MPs. In political science (see Osborn, 2012), topics related to women are considered those that are, on the one hand, traditionally believed to be in the domain of women (for example, education, healthcare, etc.), and on the other that concern women’s well-being directly (for example, childcare, domestic violence, equality, etc.). Similarly, Bäck et al. (2014), Hansen et al. (2018) and Mensah and Wood (2018) found in the corpora of parliamentary speeches from Sweden, Denmark and Ghana, respectively, that women more often spoke about soft policy areas in comparison to men who more often tackled hard policy areas. This terminology is adopted from political science (e.g., Wängnerud, 1996), where policy areas are divided into so-called hard (e.g., Macroeconomics, Energy, Transportation, 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, 1997; Litosseliti, 2006). Furthermore, as Bing and Bergvall (1998) point out, similarities in language use of different genders are often overlooked despite being more outstanding than the differences. Likewise, Blaxill and Beelen (2016) have shown a similar tendency in the context of parliamentary discourse research. Therefore, 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 and Patterson, 2015).

# Corpus analysis

In this section, we will explore a large corpus of Slovenian parliamentary debates. We will demonstrate how basic corpus analysis techniques (for more, see 3 Corpora and concordancers) 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 topics related to women have been debated by female and male speakers in the past 25 years*. We will first learn **how to use and compare relative frequencies** in subcorpora of different sizes. Then, we will learn **how to extract collocations** of selected nouns in the subcorpora.

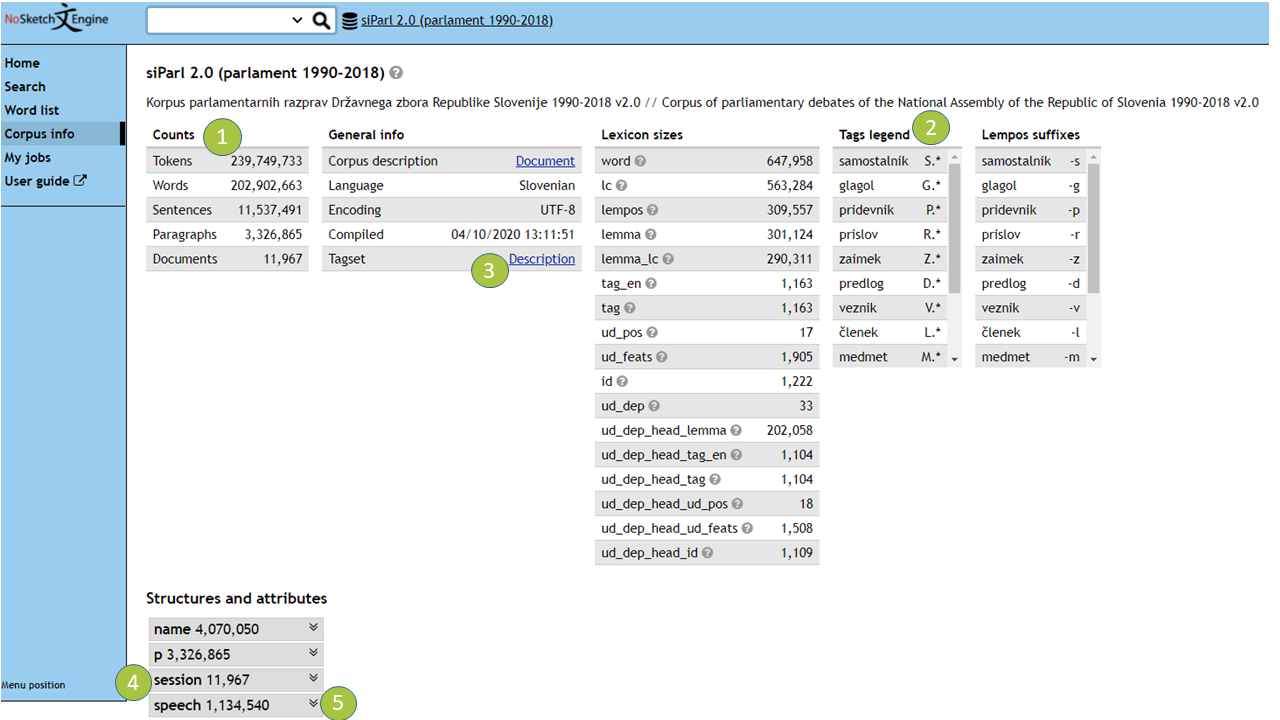
## The siParl 2.0 corpus

The [siParl corpus](http://hdl.handle.net/11356/1300) 2.0 (Pančur et al., 2020) is composed of parliamentary debates of the [National Assembly](https://en.wikipedia.org/wiki/National_Assembly_(Slovenia)) of the [Republic of Slovenia](https://en.wikipedia.org/wiki/Slovenia) from 1990 to 2018. The corpus contains records of parliamentary debates from the [period of secession](https://en.wikipedia.org/wiki/Slovenia#Slovenian_Spring,_democracy_and_independence) from [Yugoslavia](https://en.wikipedia.org/wiki/Breakup_of_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 terms, i.e. 1996-2018, and minutes of the Council of the President of the National Assembly from the second to the seventh parliamentary terms, 1996-2018.

The corpus contains metadata about the speakers (e.g., speaker’s name, gender, role, party), a typology of parliamentary sessions (e.g., regular/urgent/ceremonial meeting) and structural and editorial annotations (e.g., legislative period, session year/title, [named entity](https://en.wikipedia.org/wiki/Named_entity)). The corpus is linguistically annotated for part-of-speech and lemma. It also contains some other annotation layers (e.g., Universal Dependencies, syntactic parsing) which will not be used in this tutorial. The corpus comprises over a million speeches or 200 million words delivered by almost 8,500 speakers (e.g., Members of Parliament, members of the government, ministry representatives, representatives of professional organizations, non-governmental organizations, and interest groups). For general information about corpora and parliamentary records see sections 3 Corpora and concordancers and 4 Parliamentary records.

This information is also summarized [in the concordancer](https://www.clarin.si/noske/run.cgi/corp_info?corpname=siparl20&struct_attr_stats=1) (see Figures 3a and 3b):

* 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. The *session* level represents a single session (see *Item 4*) and conveys information such as session date, type and title (see Figure 3b).
* In addition, structural attributes at the *speech* level represent one utterance of a speaker (see *Item 5*) and give information on speaker’s name, gender, role and party affiliation (see Figure 3b).

****

**Figure 3a: The siParl 2.0 information page.**

****

**Figure 3b: The siParl 2.0 information page with extended section “Structures and attributes”.**

## TASK 1: Representation of women in the Slovenian Parliament

The Slovenian National Assembly has 90 MPs, including one representative for the Italian and one for the Hungarian minorities, who are members of [nine political parties](https://en.wikipedia.org/wiki/List_of_political_parties_in_Slovenia). Slovenia is one of the youngest EU member states and has seen dramatic changes in gender equality in the past 30 years. While in 1986 when Slovenia was still part of the Federal Socialist Republic of Yugoslavia, female MPs represented almost a quarter of all MPs, in 1992 when the National Assembly of the Republic of Slovenia was convened for the first time, this share notably significantly, so that only in ten MPs was female (Selišnik and Antić Gaber, 2017). During the transition process, when social, political, economic and value systems changed fundamentally, 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 (1/8/2014-21/6/2018), which is the last one included in the siParl 2.0 corpus, female MPs held just over 40% of the seats, largely due to legislative measures enforcing gender quotas.[[1]](#footnote-1) According to the [EU gender equality value and political power index for 2017](https://eige.europa.eu/gender-equality-index/2017/compare-countries/index/bar), Slovenia ranked among the top 10 EU countries in the proportion of women MPs.

In Task 1, we are interested in comparing the contents of the siParl 2.0 corpus with the trends observed in parliamentary elections and the Slovenian society.

### Creating subcorpora

By taking advantage of the metadata available in the corpus (see Figure 3b), we split the corpus into parts, called [subcorpora](https://www.sketchengine.eu/guide/create-a-subcorpus/), according to the following criteria:

* **Gender of the speaker.** Each speaker in the corpus is labelled with one of the following gender categories based on the data provided by the official records of the National Assembly*: male, female* or *unknown* (in cases when the metadata records are incomplete). In this tutorial we only use the first two categories, male and female, and disregard the speakers with unknown gender.
* **Role of the speaker.** Each speaker in the corpus is attributed one of the two roles: *a Member of Parliament (MP)* or *external speaker.* The latter are invited speakers who are not members of the Parliament, such as members of the government along with representatives of ministries and non-governmental organizations. In this tutorial, we only use the category of MPs since we wanted to study only the contributions by the elected representatives of the Slovenian people.
* **Type of the speaker.** Apart from the role, each speaker in the corpus is also labelled with one of the following two speaker types*: chairperson* or *regular speaker*. Regular speakers are all the speakers in the Parliament who have been explicitly given the floor by the chairperson in charge of the meeting. In this tutorial we only use the category of regular speakers. We intentionally exclude the chairpersons 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 in our particular task.
* **Parliamentary term.** Each speech in the corpus is categorized into the legislative period in which it was delivered (Term 1 – Term 7). Since it is out of scope for our study, we exclude from our analysis the [period of secession](https://en.wikipedia.org/wiki/Slovenia#Slovenian_Spring,_democracy_and_independence) from [Yugoslavia](https://en.wikipedia.org/wiki/Breakup_of_Yugoslavia) which is labelled as Term11 in the corpus.
* **Working body.** The siParl 2.0 corpus contains speeches from plenary sessions, working bodies of the National Assembly (e.g., The Commission for Petitions, Human Rights and Equal Opportunities) and from the Council of the President of the National Assembly.In this tutorial we only analyse the speeches delivered at plenary sessions.

Based on these criteria, we first create a total of 14 subcorpora, one for each of the seven parliamentary terms containing the speeches by female MPs and one for male MPs, in their role as regular speakers. In addition to these 14 subcorpora, we also create subcorpora that contain all speeches by male and female MPs for each term separately, and for all of the seven terms together.

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

The created subcorpora and information on their size are available [here](https://www.clarin.si/noske/run.cgi/subcorp?corpname=siparl20).

For advanced users of the tutorial we also provide an example of the CQL commands that are used to generate the following subcorpus:

**Term1-Female[[2]](#footnote-2):**

<speech gender="F" & role\_en="MP" & type\_en="Regular speaker"/> within <session mandate\_en="Term 1 .\*" & organ="Državni zbor Republike Slovenije"/>

This command searches for all utterances (attribute *speech*) spoken by speakers whose gender is female(attribute *F*), whose role is member of parliament (attribute MP) and whose type is regular speaker *(*attribute *Regular speaker*) within the texts from the first parliamentary term (attribute *Term 1 .\**) that were given in the *National Assembly of the Republic of Slovenia* (attribute (in Slovene): *Državni zbor Republike Slovenije*).

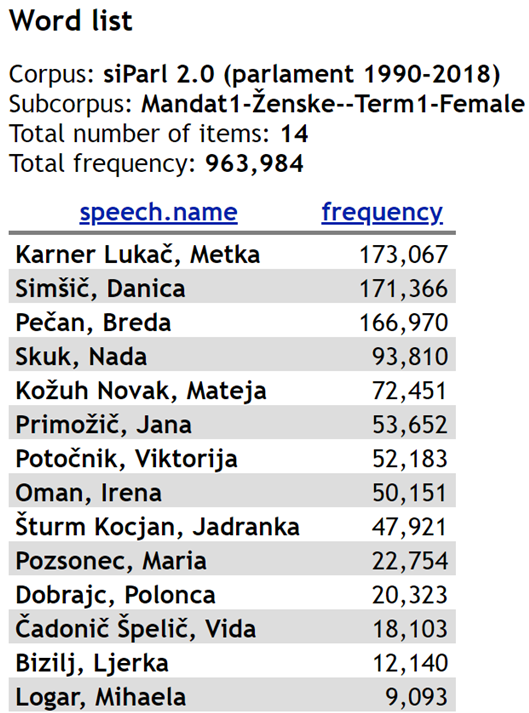
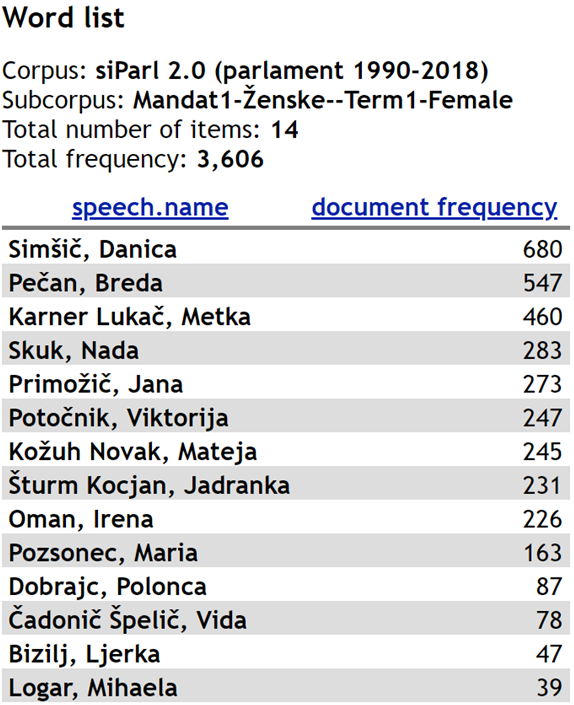
In order to compile a subcorpus of speeches delivered by male MPs, we only need to change the *speech gender* attribute in the CQL command above (i.e., *speech gender="M"*). Similarly, if we want a subcorpus containing all speeches from all terms delivered by both female and male MPs, we first have to include both genders through the *speech gender* attribute (i.e., *speech gender="F|M"*) and all terms through the *session mandate* attribute (i.e., *session mandate\_en="Term .\*"*).

### Using frequency lists

Using the subcorpora created in the previous step, we will analyse the contribution of the female and male MPs across time with the help of one of the most basic corpus techniques, called [frequency lists](https://www.sketchengine.eu/guide/wordlist-frequency-lists/#toggle-id-2-closed). 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 this function to obtain information on the number of male and female MPs in each parliamentary term, and the number of speeches and tokens they uttered.

A screencast of how to create a frequency list in NoSketchEngine is available here.

An example of two frequency lists can be seen in Figure 4.



**Figure 4. Frequency lists for the female MPs from Term1.** [**On the left**](https://www.clarin.si/noske/run.cgi/wordlist?corpname=siparl20&attrs=word&ctxattrs=word&structs=session%2Cspeech%2Cgap%2Cg&refs=%3Dspeech.name&pagesize=50&wlmaxitems=100&wlsort=f&subcnorm=freq&corpname=siparl20&reload=&usesubcorp=Mandat1-%C5%BDenske--Term1-Female&wlattr=speech.name&usengrams=0&ngrams_n=2&ngrams_max_n=2&nest_ngrams=0&wlpat=&wlminfreq=1&wlmaxfreq=0&wlfile=&wlblacklist=&include_nonwords=1&wlnums=docf&wltype=simple) **we show the number of speeches they have made, and** [**on the right**](https://www.clarin.si/noske/run.cgi/wordlist?corpname=siparl20&attrs=word&ctxattrs=word&structs=session%2Cspeech%2Cgap%2Cg&refs=%3Dspeech.name&pagesize=50&wlmaxitems=100&wlsort=f&subcnorm=freq&corpname=siparl20&reload=&usesubcorp=Mandat1-%C5%BDenske--Term1-Female&wlattr=speech.name&usengrams=0&ngrams_n=2&ngrams_max_n=2&nest_ngrams=0&wlpat=&wlminfreq=1&wlmaxfreq=0&wlfile=&wlblacklist=&include_nonwords=1&wlnums=frq&wltype=simple) **the number of tokens in these speeches.**

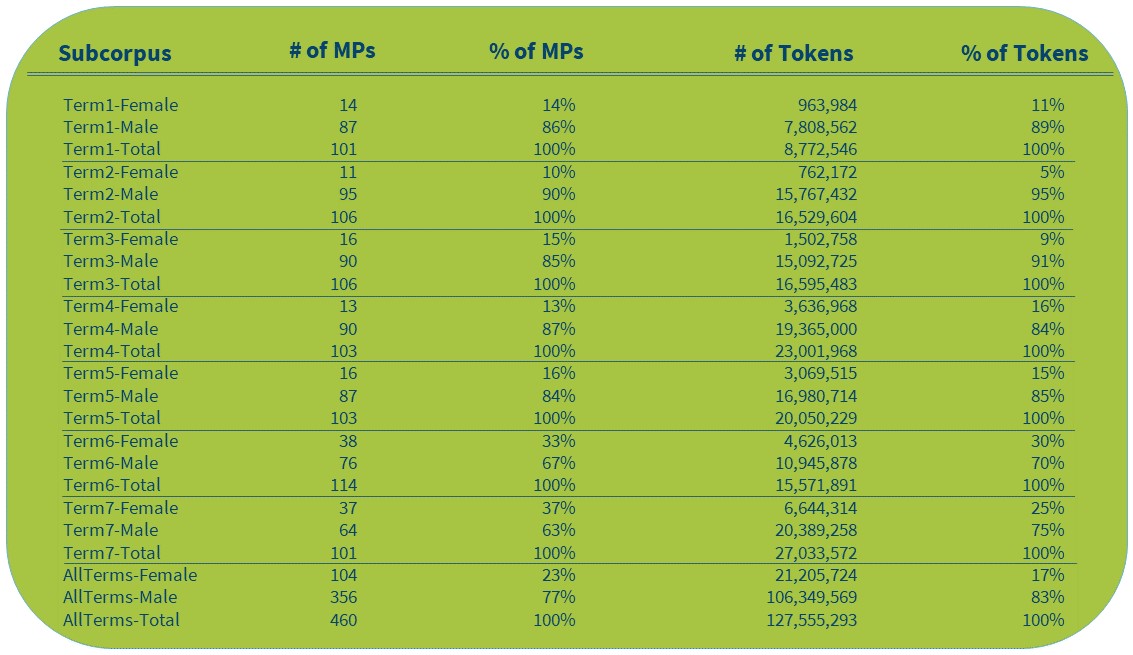
As can be seen from Figure 4, there were a total of 14 female MPs in Term1 who collectively uttered more than 3,500 speeches, with just under one million tokens. On average, this is almost 260 speeches or 69,000 tokens per MP. However, the contribution per MP is very uneven, ranging from nearly 700 speeches, or more than 170,000 tokens, to fewer than 40 speeches comprising around 9,000 tokens.

Measured in speeches, the first-ranking speaker is Danica Simšič, a member of the Democratic Party of Slovenia, a small opposition party who was elected to the Parliament only in the first term and contributed 680 speeches, or almost 19% of all the speeches in the entire subcorpus, which is almost three times the female average. The share of the speeches by Danica Simšič was calculated by dividing the number of her speeches by the number of all the speeches by female MPs (see Figure 4, left: *Total frequency* and *Simšič, Danica*). The ratio with regard to the average was calculated by dividing the number of speeches by Danica Simšič with the average number of all speeches by female MPs, which we get by dividing the total number of speeches by female MPs with the number of all female MPs (see Figure 4, left: *Total frequency* and *total number of items*).

Measured in tokens, the female MP with the largest contribution in this subcorpus is Metka Karner Lukač, member of the Slovenian People’s Party, the oldest party in Slovenia, who gave speeches with 173,067 tokens or 18% of the entire subcorpus, nearly three times the average. The MP with the lowest number of speeches as well as tokens is Mihaela Logar, a member of the Slovenian People’s Party, who spoke 39 times and gave speeches with just over 9,000 tokens in her four-year term as MP, which is nearly 20 times less than the first-ranking speaker. It is interesting to note that approximately half of all contributions can be attributed to only three MPs: the two mentioned above and to Breda Pečan, member of the Social Democrats, whose speeches on average were also the longest. This information can be extracted from data on the average length of speeches for individual MPs, which we get by dividing the total number of tokens for the individual MP (see Figure 4, right) with the total number of speeches by this particular MP (see Figure 4, left).

### Comparative analysis

For a comparative analysis of the representation of men and women in the Slovenian Parliament over time, we record the number of MPs and the number of tokens in each of the 14 subcorpora (see section 6.2.1 [Creating subcorpora](#_Creating_subcorpora)). Since the concordancer does not provide such summary tables, we manually enter the values into a spreadsheet, as can be seen in Table 1. Visualization of the results can be seen in Figure 5 and Figure 6.

****

**Table 1. Overview of the sizes of the siParl 2.0 subcorpora.[[3]](#footnote-3)**

**Figure 5: Share of male and female MPs in siParl 2.0 over time.**

**Figure 6: Contribution of male and female MPs in siParl 2.0 measured in uttered tokens over time.**

The results show that the share of tokens in speeches by female MPs gradually increases along with the number of female MPs in the Parliament. A notable leap in the representation of women can be observed in the last two terms covered by siParl 2.0, which is most probably due to the adoption of gender quotas, according to which, since 2011, 35% of the candidates must be women (Selišnik and Antić Gaber, 2017). Importantly, despite quotas being one of the most successful measures to increase the number of female MPs (Selišnik and Antić Gaber, 2017), our corpus analysis shows a large discrepancy between the number of female speakers and the amount of tokens in their speeches. While election results are easily accessible and frequently used by researchers and analysts, the amount and content of the speeches has been much less readily available and studied much less comprehensively outside corpus linguistics, which is a gap we aim to help close with this tutorial. By looking into the created subcorpora, we can see that for female MPs there is a clear tendency to contribute a lower share of content than would be expected given their share of speakers. This is especially true for Term2 where they produced two times fewer tokens than would correspond to their speaker share. In fact, only in Term4 was women’s contribution marginally greater than their share of seats in the Parliament. Therefore, we can conclude that having more elected female MPs does not directly mean that their voice is more present in the National Assembly. Nonetheless, as the number of female speakers in the Parliament has been on the rise in recent years, the number of their speeches has started increasing as well. From as little as a tenth of all the speeches given in the Parliament during the first five parliamentary terms by female MPs, the contribution of female MPs has risen to almost a third in the last two parliamentary terms covered by our corpus.

## TASK 2: Issues addressed by women

Studies of female political discourse have shown that women tend to debate different topics to men. More about the impact of gender on language can be found in the section 5 Language and gender.

In Task 2, we are interested in comparing the topics discussed by female MPs in siParl 2.0 with those discussed by their male counterparts. While topic classification is readily available for some parliamentary corpora, as in Nanni et al. (2018), this is not the case for siParl 2.0. Topic classification could be derived automatically, as in Karan et al. (2016), however, the goal of this tutorial is to demonstrate the potential of parliamentary corpora research via concordancers without requiring programming skills, which is why we opt 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. It should be noted that both automatic and manual annotation approaches produce errors, but that we can avoid those due to annotator’s subjectivity by employing multiple annotators. The final quality of such manual annotations can be verified with statistical tests that report on the [inter-annotator agreement](https://en.wikipedia.org/wiki/Inter-rater_reliability).

### Extracting keywords

To enable the comparative analysis we will first use a well-known corpus analysis technique called [keyword extraction](https://en.wikipedia.org/wiki/Keyword_(linguistics)). 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 2.0 subcorpora to uncover the most prominent issues discussed by the MPs in the Slovenian Parliament.

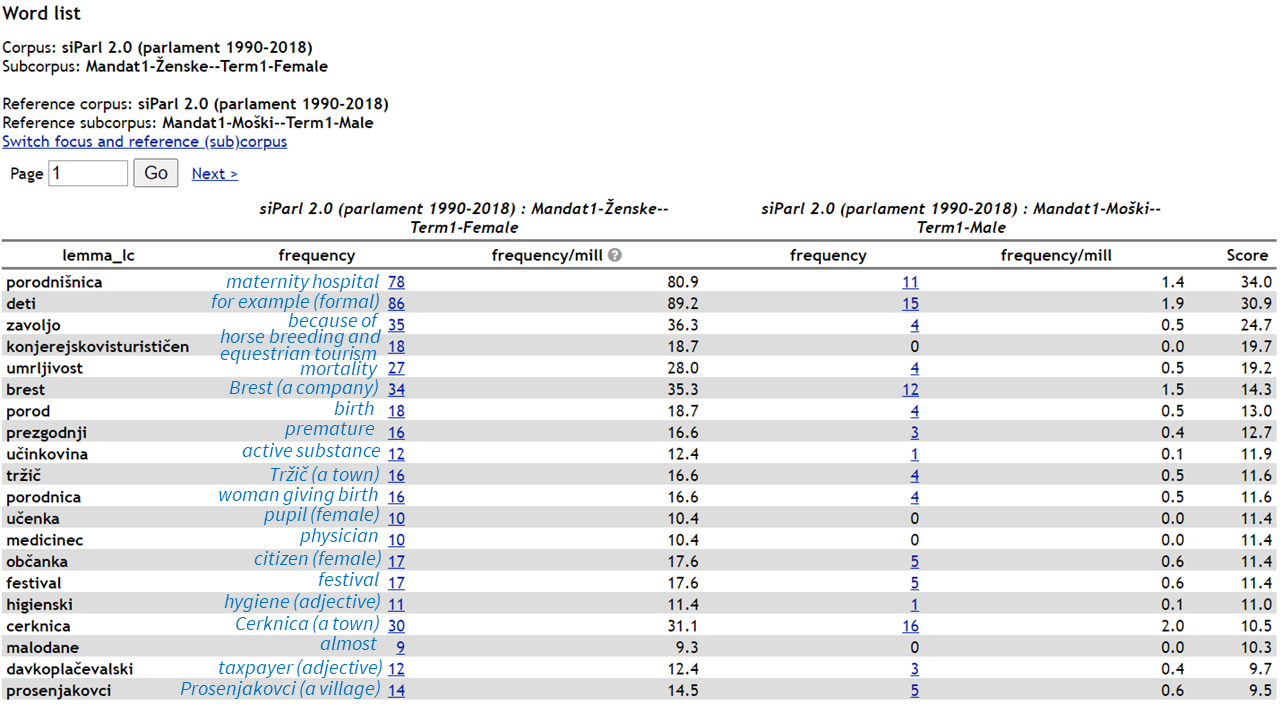
A screencast of how to generate a keyword list in NoSketchEngine is available here.

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

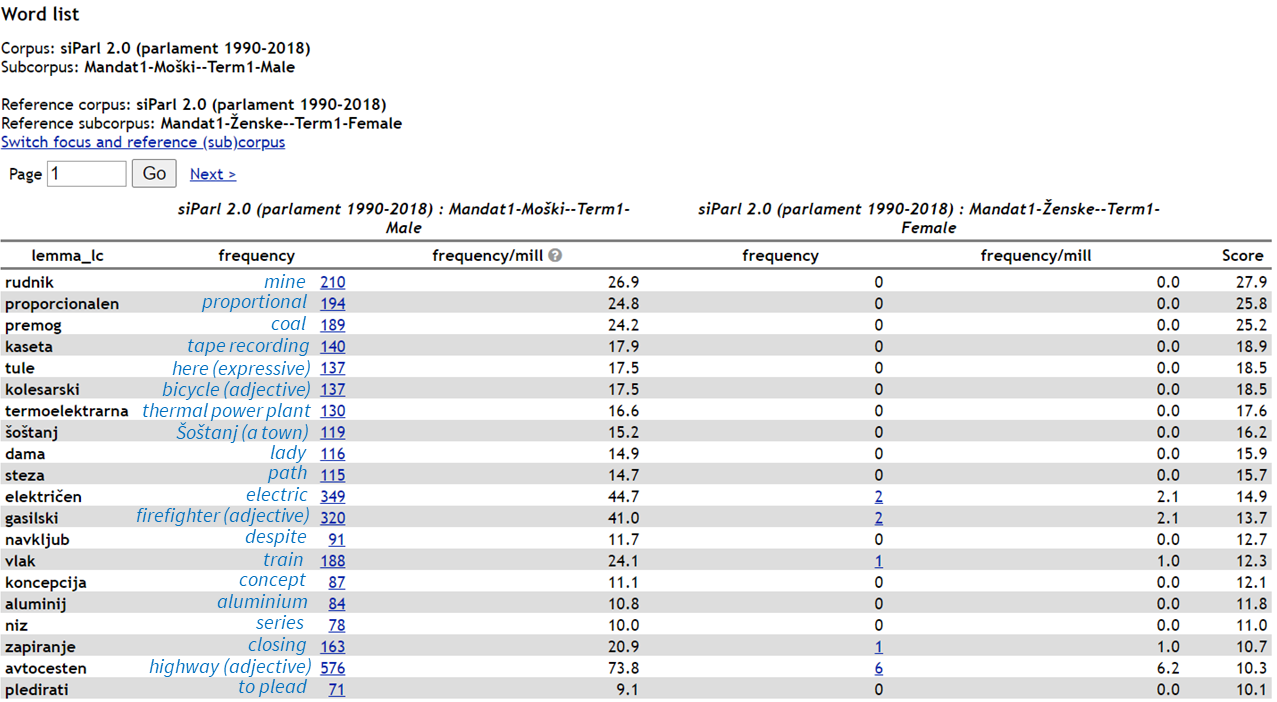
* [Term1-Female](https://www.clarin.si/noske/run.cgi/wordlist?corpname=siparl20&attrs=word&ctxattrs=word&structs=session%2Cspeech%2Cgap%2Cg&refs=%3Dspeech.name&pagesize=50&wlmaxitems=100&wlsort=f&subcnorm=freq&corpname=siparl20&reload=&usesubcorp=Mandat1-%C5%BDenske--Term1-Female&wlattr=lemma_lc&usengrams=0&ngrams_n=2&ngrams_max_n=2&nest_ngrams=0&wlpat=&wlminfreq=5&wlmaxfreq=0&wlfile=&wlblacklist=&wlnums=frq&wltype=keywords&ref_corpname=siparl20&ref_usesubcorp=Mandat1-Mo%C5%A1ki--Term1-Male&simple_n=1)
* [Term1-Male](https://www.clarin.si/noske/run.cgi/wordlist?;corpname=siparl20;usesubcorp=Mandat1-Mo%C5%A1ki--Term1-Male;wlattr=lemma_lc;wlminfreq=5;wlmaxfreq=0;wlpat=.%2A;wlmaxitems=100;wlsort=f;ref_corpname=siparl20;ref_usesubcorp=Mandat1-%C5%BDenske--Term1-Female;wlcache=;simple_n=1.0;wltype=keywords;wlnums=frq;include_nonwords=0;blcache=;wlpage=1;usengrams=0;ngrams_n=2;ngrams_max_n=2;nest_ngrams=0;complement_subc=0)
* [Term7-Female](https://www.clarin.si/noske/run.cgi/wordlist?corpname=siparl20&attrs=word&ctxattrs=word&structs=session%2Cspeech%2Cgap%2Cg&refs=%3Dspeech.name&pagesize=50&wlmaxitems=100&wlsort=f&subcnorm=freq&corpname=siparl20&reload=&usesubcorp=Mandat7-%C5%BDenske--Term7-Female&wlattr=lemma_lc&usengrams=0&ngrams_n=2&ngrams_max_n=2&nest_ngrams=0&wlpat=&wlminfreq=5&wlmaxfreq=0&wlfile=&wlblacklist=&wlnums=frq&wltype=keywords&ref_corpname=siparl20&ref_usesubcorp=Mandat7-Mo%C5%A1ki--Term7-Male&simple_n=1)
* [Term7-Male](https://www.clarin.si/noske/run.cgi/wordlist?;corpname=siparl20;usesubcorp=Mandat7-Mo%C5%A1ki--Term7-Male;wlattr=lemma_lc;wlminfreq=5;wlmaxfreq=0;wlpat=.%2A;wlmaxitems=100;wlsort=f;ref_corpname=siparl20;ref_usesubcorp=Mandat7-%C5%BDenske--Term7-Female;wlcache=;simple_n=1.0;wltype=keywords;wlnums=frq;include_nonwords=0;blcache=;wlpage=1;usengrams=0;ngrams_n=2;ngrams_max_n=2;nest_ngrams=0;complement_subc=0)

We first export the keyword lists for each of the four subcorpora into a spreadsheet and then manually annotate the data (see section 6.3.3 Comparative analysis).

Figures 7 and 8 show the 20 top-ranking keywords for female and for male MPs in Term7, using the [Simple Maths statistics](https://www.sketchengine.eu/documentation/simple-maths/). Stark differences can be observed: while the majority of all the displayed keywords for female speakers are related to Healthcare, the key male vocabulary belongs to the domains of *Foreign affairs*, *Infrastructure* and *Justice*.



**Figure 7: Twenty top-ranking keywords of female MPs (with their English translations in blue) in Term1 compared to those of their male counterparts.**

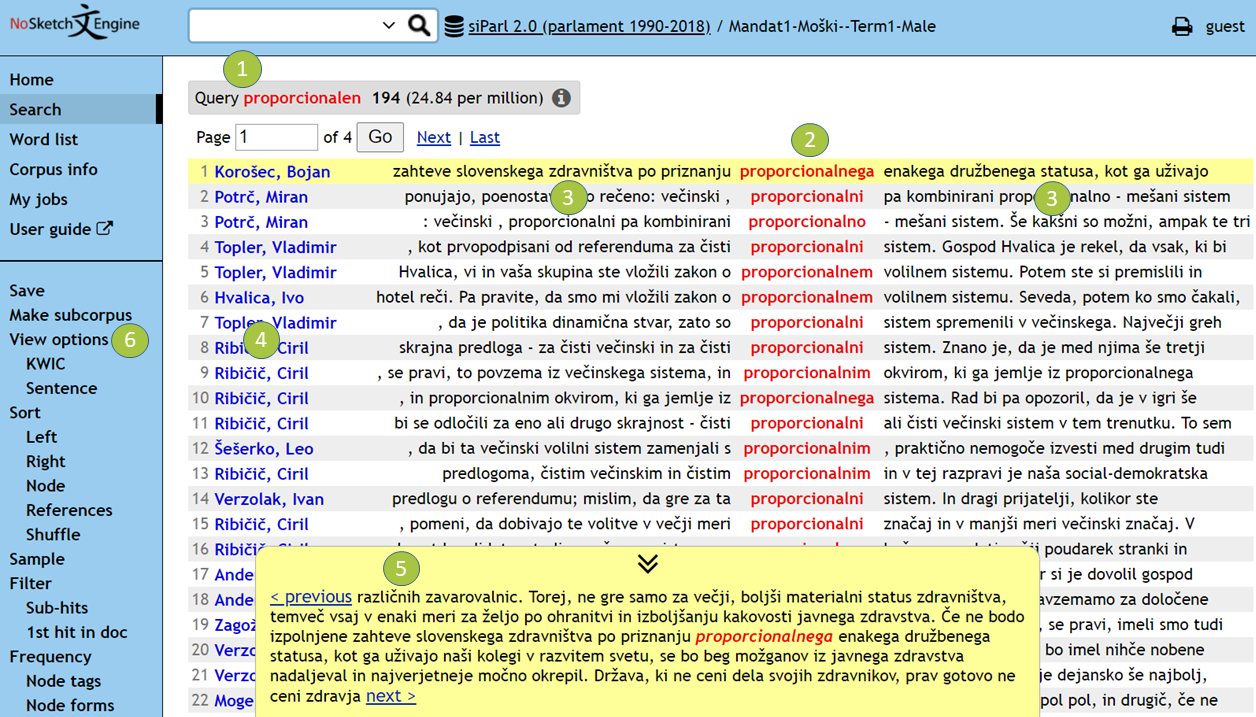


**Figure 8: Twenty top-ranking keywords of male MPs (with their English translations in blue) in Term1 compared to those of their female counterparts.**

### Analysing concordances

For further analysis, we select the 100 top-ranking key lemmas, excluding personal names, from the four generated keyword lists. Our goal is to manually categorize each of them into topics after inspecting their [concordances](https://www.sketchengine.eu/guide/concordance-a-tool-to-search-a-corpus/), which are lists of all instances of the search words in their contexts, as shown in Figure 9:

* Concordances can be displayed directly by clicking on the key lemma in the keyword list. At the top of the page that appears, the selected lemma is displayed along with its hits (see *Item 1*).
* 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.
* Alternatively, you can also tweak the view parameter in *View options* (see *Item 6*) and display more metadata or a wider context.

****

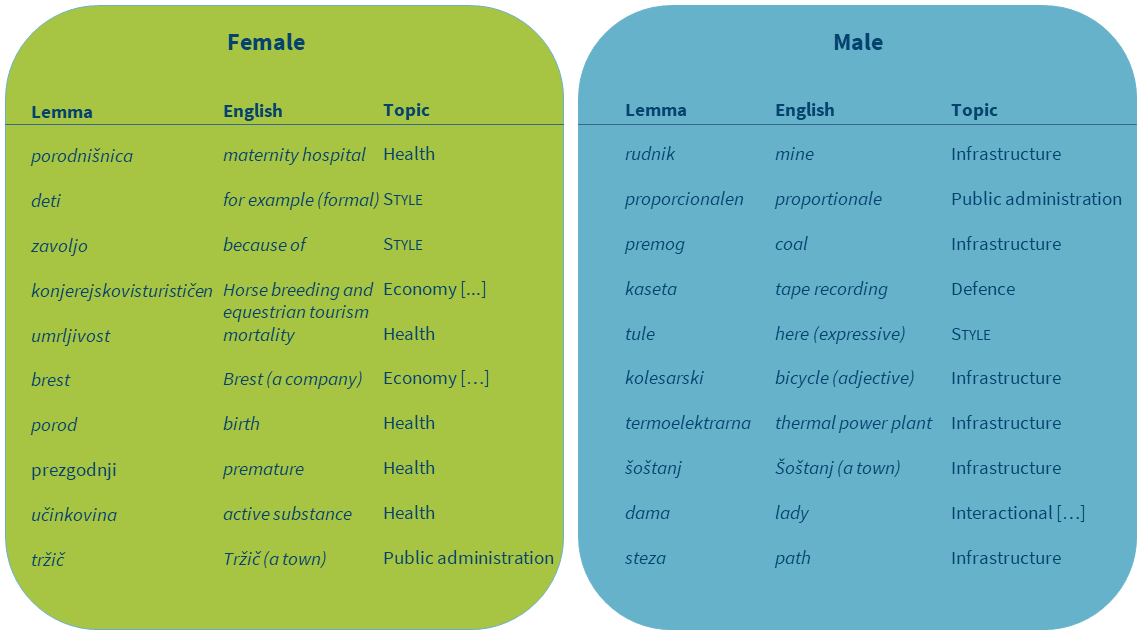
**Figure 9: Extended context for the first hit in the** [**concordance list**](https://www.clarin.si/noske/run.cgi/view?corpname=siparl20;usesubcorp=Mandat1-Mo%C5%A1ki--Term1-Male;q=q%5Blemma_lc%3D%3D%22proporcionalen%22%5D;complement_subc=0) **of the keyword “proporcionalen/proportional”.**

### 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 will use the list of [14 ministries of the current Slovenian government](https://www.gov.si/en/state-authorities/ministries/) as categories for topic analysis. The categories are listed in Table 2. While several other lists of topics could be used, this one seems the most natural in the specific setting of parliamentary discourse. In addition to these 14 categories, the list is extended with another fourcategories for keywords that cannot be listed elsewhere: *Multiple* (for keywords used in discussions on several topics); *style* (for clearly colloquial or jargon keywords used only by specific speakers); *ideology* (for keywords used for ideological labelling, e.g., “partijski/partisan (derogatory, adjective)”); and *interactional/procedural* (for keywords referring to other MPs or to procedural matters). Illustrative examples of manual annotation of the 10 top-ranking keywords by female and by male MPs are given in Table 3.

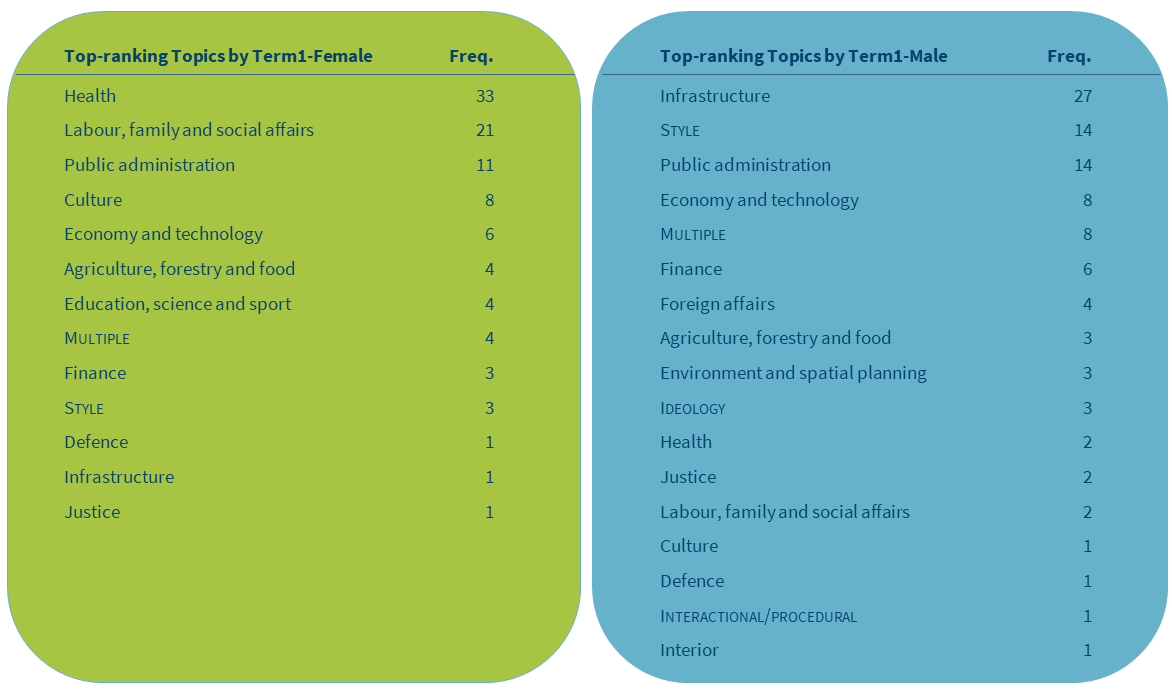
****

**Table 2. Keyword annotation categories. We use upper case for the last four function categories in order to separate them from content categories.**

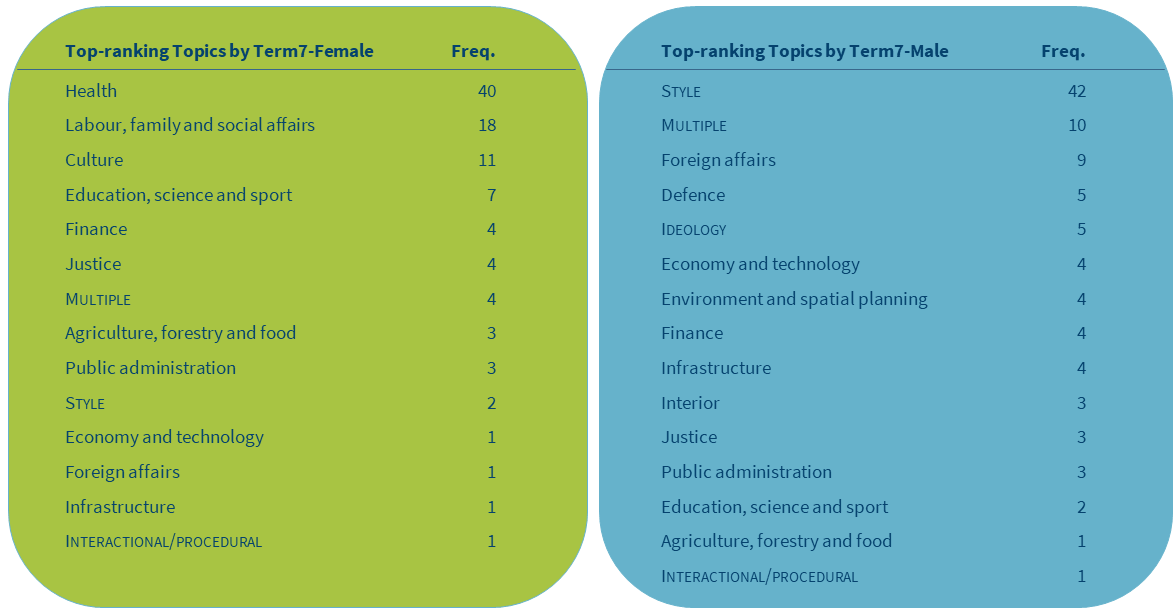
****

**Table 3: Illustrative example of manual topic annotation of 10 top-ranking keywords from speeches by female and male MPs in Term1.**

Tables 4 and 5 contain summarized results of the manual annotation of 100 top-ranking key lemmas for female and male MPs in Term1 and Term7, respectively. 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 one.

****

**Table 4: Topics of the 100 top-ranking keywords in speeches by female and male MPs in Term1.**



**Table 5: Topics of the 100 top-ranking keywords in speeches by female and male MPs in Term7.**

In Term1, the majority (54%) of all the analysed female keywords belong to only two topics (*Health* and *Labour, family and social affairs*), whereas the two most common topics of male MPs with a similar share (55%) are *Infrastructure* and *Public administration*. In Term7, the two prevailing female topics not only remained the same but also slightly intensified (58%) while for male MPs the most prominent topic became *Foreign affairs.* While the division of topics between genders cannot be explained through simple corpus analysis, both an intense focus on *Health* and *Social affairs* in the female subcorpora, as well as the shift from *Infrastructure* and *Public administration* to *Foreign affairs* in the male subcorpora are a good reflection of the state of affairs. At the time of independence, the new-born state of Slovenia had to create its own public administration and re-build its infrastructure, which required intensive discussions in the first parliamentary term. The last term, on the other hand, was marked by more intensive international trade as well as greater international security threats, which warranted legislative and budgetary decisions in the parliament. The intensified discussions on *Health* *issues* as well as a continued focus on *Social issues* in Term7 were largely due to heavy pressure on the budget due to the severe economic crisis in that period, which also negatively impacted the already crumbling public health system.

We can also observe that in Term1, men and women attached similar importance (according to the share of the analysed keywords) to more topics than in Term7 (five shared topics with equal importance in Term1 vs. three topics in Term7). Thus, in Term1, men debated slightly more than women on *Public administration* (M: 14% vs. F: 11%), *Economy and technology* (M: 8% vs. F: 6%) and *Justice* (M: 2% vs. F: 1%), and women debated slightly more on *Agriculture, forestry and food* (M: 3%, F: 4%), while MPs of both genders debated equally on *Defence* (1%). In Term7, on the other hand, only three topics appear as equally important to male and female MPs: *Finance* (4%), *Justice* (3%) and *Public administration* (3%). For a more in-depth analysis, it may be interesting to look into the specific ways (if any) in which female and male MPs approach and discuss the key issues of these topics.

The data shows that female-specific topics are *Education* in Term1, and *Health*, *Labour, family and social affairs* and *Culture* in Term7. It is surprising that these last three topics, which are the top three for women, do not appear on the list of topics for men, while we can find the top topics for men (*Infrastructure, Public administration, Foreign affairs*) on the women’s list (although with a negligible share of the total). Male-specific topics, on the other hand, are *Environment* and *Interior* in both parliamentary terms, *Foreign affairs* (in Term1) and *Defence* (in Term7). Surprisingly, it can be observed that *Economy and technology*,which is a topic present in both terms for both genders, drastically loses its saliency in the subcorpus of female speeches in Term7.

Apart from the different areas of interest, the results also show that male and female MPs have different debating styles. In both terms, male MPs use considerably more stylistic words (i.e., clearly colloquial or jargon words observed only from specific speakers) than their female counterparts. In fact, the usage of such words by male MPs tripled in Term7 compared to Term1, which on the one hand indicates the greater casualness of male MPs in general, as well as highlights a change in the debating culture, which in the last two decades has become more ludic and informal. We also observe that ideological words, which tend to be polarizing, are only used by the male MPs. This is in line with findings from the literature, according to which men’s conversational strategies are more aggressive and competitive, whereas women’s debating style is more cooperative (Coates, 1997). It should be noted that these results would need to be confirmed by a more detailed analysis, since some recent studies have challenged the clear-cut gender difference in debating styles as being an overgeneralization (cf. Ilie, 2013).

Our findings point out notable differences in the roles and interests of male and female MPs which is in line with previous studies showing that women focus more on the so-called soft topics than men. Diachronic comparisons reveal that *Health* and *Social issues* remain the top priority for female MPs, while the focus of male MPs shifted from *Infrastructure* and *Public administration* towards *Foreign affairs.* In Term1, the most prominent common topic was *Public administration*, while in Term7, *Finance* became the most often discussed topic for both male and female MPs. This, however, does not mean that female MPs now debate more on the hard topics, quite the contrary, in fact, as our analysis showed that in Term7 women focused more on soft topics than in Term1.

## TASK 3: Topics related to women

This task is inspired by related work (see Blaxill and Beelen, 2016) which investigated how frequently, by whom and in what way are topics related to women, such as women’s rights, equality, discrimination, etc., 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 parliament, such as Sweden, for which Bäck et al. (2014) found that female MPs discuss the so-called hard policy issues less frequently. Furthermore, Antić Gaber and Ilonszki (2003) state that society usually expects women MPs to be actively involved in different policy areas than men, in those that are particularly salient to women because of women’s historic role in society or because those areas directly affect women’s lives.

Given that debates about gender equality are still dominated by issues pertaining to women’s unequal and special role in society, which directly impacts their rights as well as discriminatory practices, we are interested in Task 3 in comparing how male and female MPs in the Slovenian Parliament express themselves when addressing topics related to women by focusing on their use of the noun “ženska/female” as an explicit indicator of debate on such topics.

### Working with frequencies

First, we are interested in how frequently the lemma of the noun “ženska/female” is mentioned by male and female MPs in different time periods (1992-2018). We query all the subcorpora and record all the frequency counts of the lemma in a spreadsheet, with the results shown in Table 6. Because we are interested in comparing subcorpora of different sizes, it is important to use [normalized frequencies](http://www.thegrammarlab.com/?p=160) instead of the absolute ones, as these raw frequencies can be misleading. Looking for example at the raw frequencies for Term2 (F: 190; M: 794), we could conclude that male MPs mention the noun “ženska/female” four times more often than their female counterparts. But, because the female subcorpus in that parliamentary term is much smaller than the male subcorpus, the normalized frequency, which calculates the frequency based on the same number of words in each subcorpus (in our case a million words), shows that the noun is actually five times more frequent in the female subcorpus (249.29) compared to the male one (50.36).

If you would like to see how to query a subcorpus for the lemma, please watch the screencast on how to extract collocations in NoSketchEngine (min. 00:09–1:24) here.

A concordance list for the lemma of the noun “ženska/female” is available here:

* [AllTerms-Female](https://www.clarin.si/noske/run.cgi/first?corpname=siparl20&reload=&iquery=&queryselector=lemmarow&lemma=%C5%BEenska&lpos=-s&phrase=&word=&wpos=&char=&cql=&default_attr=word&fc_lemword_window_type=both&fc_lemword_wsize=5&fc_lemword=&fc_lemword_type=all&fc_pos_window_type=both&fc_pos_wsize=5&fc_pos_type=all&usesubcorp=VsiMandati-%C5%BDenske--AllTerms-Female&fsca_session.organ=&fsca_session.id=&fsca_session.title=&fsca_session.month=&fsca_session.date=&fsca_speech.who=&fsca_speech.name=)

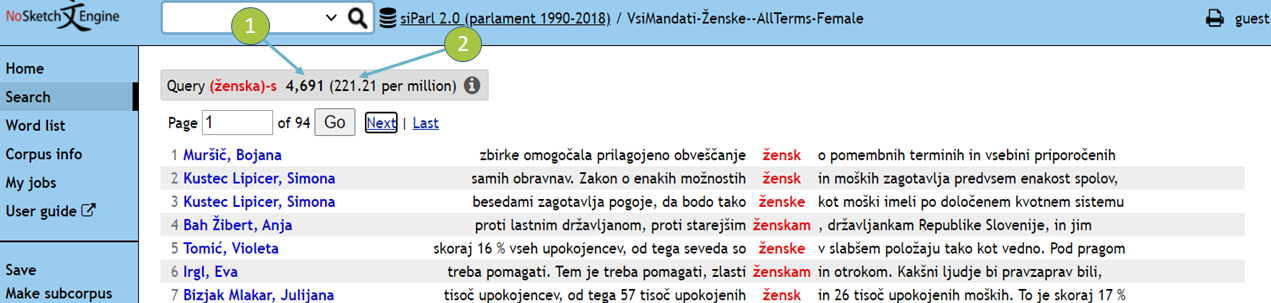
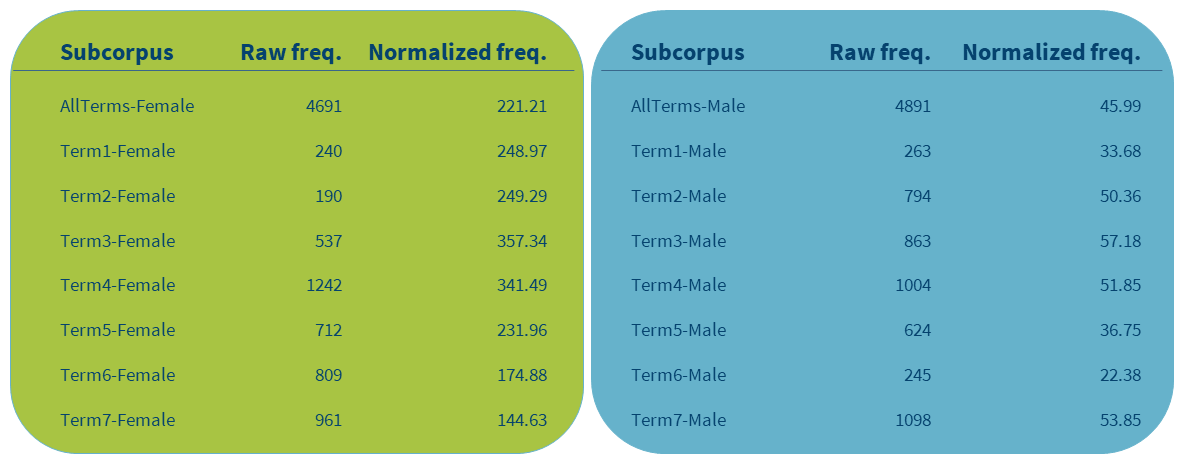


Figure 10: The beginning of the concordance list for the lemma of the noun “ženska/female” with raw frequency (*Item 1*) and normalized frequency (*Item 2*).



**Table 6. Raw and normalized frequencies of the lemma “ženska/female” in siParl 2.0.**

Overall, the normalized frequency of the word for the entire time period is nearly five times higher in the subcorpus of female MPs compared to their male counterparts (221.21 vs. 45.99). As can be seen from Figure 10, female MPs in general tend to talk more about women than their male counterparts. However, if the period between Term2 and Term4 (1996–2008) saw a strong emphasis on topics about women among female MPs in comparison to male MPs, we can observe a steep downward turn in the female subcorpus in more recent terms.

The most striking results for the female subcorpus are observed in Term3 and Term5. In Term3, the normalized frequency of the word reached a record high, which might be a consequence of lower numbers of female representatives in the previous two parliamentary terms and their unequal position in society overall. However, as the same trend is observed in the male subcorpus, it appears that other reasons had to contribute to this result as well. For a better insight into the phenomenon we perform qualitative analysis by inspecting 50 concordance lines for the five most active male MPs. They show that legislation regulating specific issues which are particularly linked to women was extensively debated at the time (i.e. legislation related to equal opportunities, artificial insemination with biomedical assistance and quotas). Almost a decade later, in Term5, there was a sudden reduction in the number of mentions of the word in the female subcorpus, which even dropped below the frequency before the year 2000 despite a higher number of female MPs. Term5 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 contextualization using qualitative methods, such as concordance analysis.

It is also interesting to observe that in Term4 despite there being the second lowest number of female MPs, the noun “ženska/female” remains very frequent as opposed to in Term6 and Term7, when the number of female MPs was considerably higher but the frequencies of the queried word almost two times lower. As previously observed in the overall production of female MPs (see 6.2.3 Comparative analysis), where we found a large discrepancy between the number of female speakers and the amount of words they uttered, these results again show that a higher number of female MPs does not guarantee a strong discussion of issues related to women in the Slovenian Parliament.

**Figure 10: Normalized frequency of mentions of “ženska/female” by female and male MPs in siParl 2.0.**

### Extracting collocations

Next, we will demonstrate another popular corpus analysis technique called [collocation extraction](https://en.wikipedia.org/wiki/Collocation), which identifies word combinations that co-occur more often than would be expected by chance based on statistical significance tests. While collocations are most typically used 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 Slovenian 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 2.0 subcorpora, one containing the female speeches from all seven parliamentary terms (*AllTerms-Female*) and the other comprising male speeches from the same time span (*AllTerms-Male*).

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

First, we extract collocations in the range one word to the left and one word to the right of the headword (i.e., the lemma of the noun “ženska/female”) 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 want to limit our analysis for this tutorial to fixed multi-word expressions, thereby using a very narrow window and strict minimum frequency criteria.

We use the [logDice](https://www.sketchengine.eu/my_keywords/logdice/) statistic measure to measure the association strength between words. While several other collocation measures are also offered by the NoSketchEngine concordancer, such as Mutual Information or T-score, we opt 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.

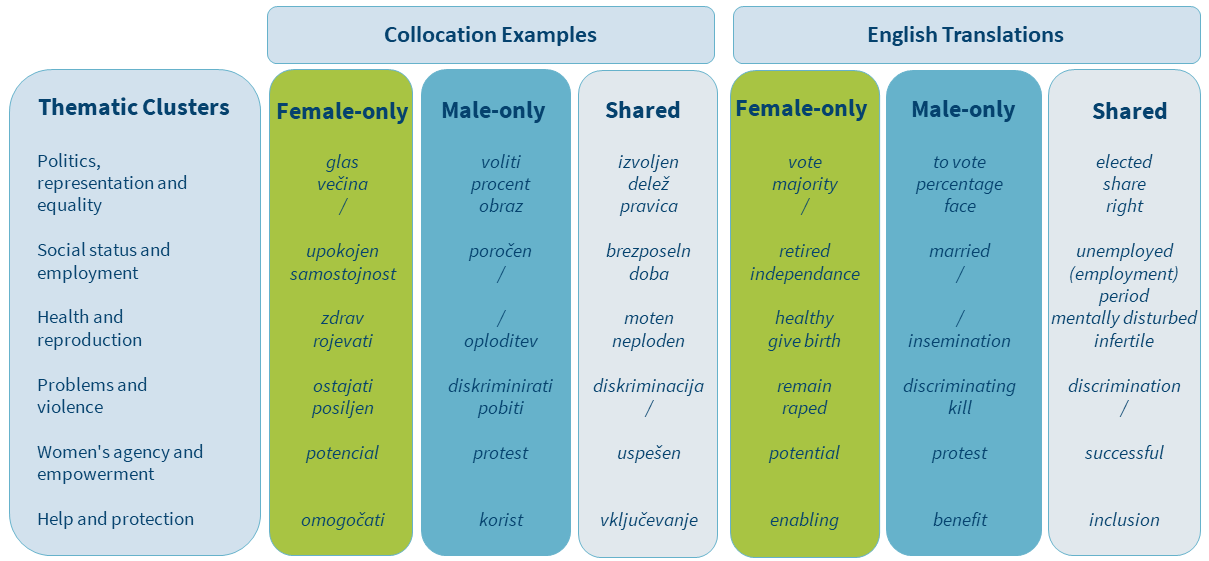
The two collocation lists for the headword “ženska/female” are available here:

* [AllTerms-Female](https://www.clarin.si/noske/run.cgi/first?corpname=siparl20&reload=&iquery=&queryselector=lemmarow&lemma=%C5%BEenska&lpos=-s&phrase=&word=&wpos=&char=&cql=&default_attr=word&fc_lemword_window_type=both&fc_lemword_wsize=5&fc_lemword=&fc_lemword_type=all&fc_pos_window_type=both&fc_pos_wsize=5&fc_pos_type=all&usesubcorp=VsiMandati-%C5%BDenske--AllTerms-Female&fsca_session.organ=&fsca_session.id=&fsca_session.title=&fsca_session.month=&fsca_session.date=&fsca_speech.who=&fsca_speech.name=)
* [AllTerms-Male](https://www.clarin.si/noske/run.cgi/first?corpname=siparl20&reload=&iquery=&queryselector=lemmarow&lemma=%C5%BEenska&lpos=-s&phrase=&word=&wpos=&char=&cql=&default_attr=word&fc_lemword_window_type=both&fc_lemword_wsize=5&fc_lemword=&fc_lemword_type=all&fc_pos_window_type=both&fc_pos_wsize=5&fc_pos_type=all&usesubcorp=VsiMandati-Mo%C5%A1ki--AllTerms-Male&fsca_session.organ=&fsca_session.id=&fsca_session.title=&fsca_session.month=&fsca_session.date=&fsca_speech.who=&fsca_speech.name=)

We then import both collocation lists into a spreadsheet and manually analyse the data.

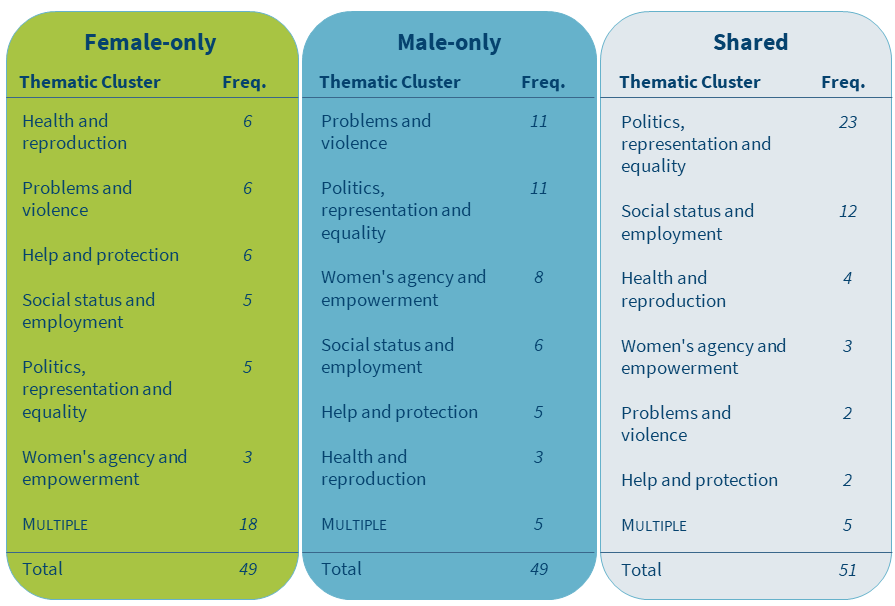
### Comparative analysis

We take 100 top-ranking collocation candidates from each list and manually divide them into three categories: female-only, male-only and shared. Next, we categorize each collocation candidate into one of seven thematic clusters that are inspired from a close reading of corpus concordances, as illustrated in Table 7. Collocation candidates that are grammatical words (such as prepositions) and are not used in concordances with a prevailing theme are assigned the label *Multiple*.



**Table 7: Examples of collocation clusters.**

The results presented in Table 8 show that just over half (51%) of the collocations are shared between MPs of both genders, revealing common ground in the understanding of women’s position in modern society. The great majority of the shared collocations (nearly 70%) fall under the first two thematic clusters: *Politics, representation and equality* and *Social status and employment*, and refer to concepts related to representation of women (e.g., “participacija/participation”), their social status (e.g., “samski/single”) and equality (e.g., “emancipacija/emancipation”).

****

**Table 8:** **Frequencies of thematic clusters for the 100 top-ranking collocation candidates in speeches by female and male MPs in all seven parliamentary terms.**

It can also be observed that the female subcorpus contains comparable representation of all but one thematic cluster which together cover almost 60% of the collocations. The only exception is the cluster *Women’s agency and empowerment*. Collocations that belong to this thematic cluster are approximately twice less frequent than collocations belonging to any other cluster. In the male subcorpus, on the other hand, the majority of collocations (61%) fall under only three thematic clusters, namely *Problems and violence, Politics, representation and equality,* and *Women’s agency and empowerment.* Furthermore, a detailed look into annotations reveals that men mainly focus on women’s social status (e.g., “poročen/married”) and problems (e.g., “zatiranje/oppression”), as well as their agency and empowerment (e.g., “sposoben/capable”), while not focusing on issues related to women’s employment and health. On the other hand, collocations from the female subcorpus first and foremost refer to health issues (e.g., “zdrav/healthy”) and measures related to help and protection (e.g., “omogočati/enabling”). These are closely followed by words referring to women’s social status (e.g., “izobrazba/education”) and violence (e.g., “posiljen/raped”), which is again contrary to the male subcorpus where more emphasis is given to problems not directly linked to violence (e.g., “poniževanje/humiliation”). In addition, similar to the male subcorpus, employment-related words do not appear in the female subcorpus among the top 100 collocation candidates.

These results indicate that female MPs try to address the multitude of issues that women face in various areas of life, from health related issues to the aggression suffered by women and difficulties with entering into politics, while male MPs focus more intensely on the unequal place of women in society and view women as active participants in the process of change. These results also confirm the findings of Antić Gaber and Ilonszki (2003) who looked at a shorter period (1996–2004) of female MPs’ activity in the Slovenian Parliament and showed with diverse, non-corpus linguistic methods that there exists a clear difference between the legislative priorities of male and female MPs, identifying similar topics to those in our analysis.

Our final observation pertains to the high number of grammatical words found in the female subcorpus, which are almost four times more frequent than in the male subcorpus. These collocation candidates in the female subcorpus mainly consists of adverbs, followed by prepositions, conjunctions, pronouns and numerals, while in the male subcorpus they predominantly include numerals. Similar to what we saw in the keyword analysis (see 6.3.3 Comparative analysis), this once more points to a different debating style of female MPs. This is in line with related work where it has been found that women tend to employ more language features such as hedges, politer forms and tag questions (Coates, 1997), more intensifiers and adjectives, especially evaluative ones (Poynton, 1989) and more personal pronouns than men, whose language includes more numbers, articles and prepositions (Newman et al., 2008).

# 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 specialized 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 also 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 maximize its potential for our research. Second, we have showcased how a set of standard corpus analysis techniques can be utilized 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 research.

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, trans-national and trans-cultural comparative research on parliamentary discourse.

# Acknowledgements

The work described in this paper was funded by CLARIN.SI – the Slovene national node of the European Common Language Resources and Technology Infrastructure CLARIN, the ARRS P6-0215 research program (Slovene language – basic, contrastive, and applied studies) and the H2020 project SSHOC – Social Sciences & Humanities Open Cloud (GA No. 823782). We would also like to thank Prof. Dr. Milica Antić Gaber, Assist. Prof. Dr. Irena Selišnik, Prof. Dr. Cornelia Ilie, Tjaša Cankar, Mladen Zobec and Ruben Ros for their thoughtful comments.

# References

Alasuutari, P., Rautalin, M., & Tyrkkö, J. (2018). 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*, *59*(3), pp. 341–363.

Antić Gaber, M., & Ilonszki, G. (2003). *Women in parliamentary politics: Hungarian and Slovene cases compared*. Ljubljana: Peace Institute, Institute for Contemporary Social and Political Studies.

Antić Gaber, M. & Selišnik, I. (2017). *The Slovene version of a “fast track” to political equality*. Teorija in Praksa, 54(2), pp. 337–354.

Bäck, Hanna., Debus, Marc., & Müller, Jochen. (2014). *Who Takes the Parliamentary Floor? The Role of Gender in Speech-making in the Swedish Riksdag.* Political Research Quarterly, 67(3), pp. 504–518.

Baker, P. (2014). *Using corpora to analyze gender*. London: Bloomsbury.

Biber, D., & Reppen, R. (2015). *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press.

Biel, Ł., Biernacka, A., & Jopek-Bosiacka, A. (2018). Collocations of Terms in EU Competition Law: A Corpus Analysis of EU English Collocations. In: Marino et al. (Eds.): *Language and Law: The Role of Language and Translation in EU Competition Law.* pp. 249–274. Cham: Springer.

Bing, J. M., & Bergvall, V. L. (1998). The question of questions: Beyond binary thinking. In J. Coates (Ed.), *Language and Gender: A Reader* (Vol. 1), pp. 496–510. Oxford: Blackwell.

Blaxill, L., & Beelen, K. (2016). A Feminized Language of Democracy? The Representation of Women at Westminster since 1945. *Twentieth Century British History*, 27(3), pp. 412–449 .

Coates, Jennifer. (1997). *Women, men and language: A sociolinguistic account of gender differences in language*. London: Longman.

Demmen, J., Jeffries, L., & Walker, B. (2018). Charting the semantics of labour relations in House of Commons debates spanning two hundred years. In: Kranert, M. and Horan, G. (eds.): *Doing Politics: Discursivity, Performativity and Mediation in Political Discourse*, *80*, pp. 81–104.

Eckert, P., & McConnell-Ginet, S. (2013). *Language and gender*. Cambridge: Cambridge University Press.

Goddard, A., & Patterson, L. M. (2015). *Language and gender*. London: Routledge.

Hansen, D. H., Navarretta, C., & Offersgaard, L. (2018). 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)*. Paris: European Language Resources Association.

Ilie, C. (2013). Gendering confrontational rhetoric: Discursive disorder in the British and Swedish parliaments. Democratization, 20(3), 501–521.

Ilie, C. (2017). Parliamentary debates. In R. Wodak & B. Forchtner (Eds.), The Routledge handbook of language and politics (pp. 309–325). Routledge.

Jaworska, S., & Ryan, K. (2018). Gender and the language of pain in chronic and terminal illness: A corpus-based discourse analysis of patients’ narratives. *Social Science & Medicine*, *215*, pp. 107–114.

Karan, M., Šnajder, J., Sirinic, D., & Glavaš, G. (2016). 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*, Berlin, Germany, pp. 12–21.

Kustec, L. S. (ur.) (2017). *Poslanke Državnega zbora: Uradno in osebno od prvega do sedmega mandata (1992-2017)*. Ljubljana: Državni zbor.

Leijenaar, M. (1997). *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.

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

Marshall, C. (2000). Policy discourse analysis: Negotiating gender equity. *Journal of Education Policy*, *15*(2), pp. 125–156.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Mensah, E. O., & Wood, S. F. (2018). 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*, *7*(2), pp. 61–77.

Mollin, S. (2007). The Hansard hazard: Gauging the accuracy of British parliamentary transcripts. *Corpora*, 2(2), pp. 187–210.

Nanni, F., Osman, M., Cheng, Y.-R., Ponzetto, S. P., & Dietz, L. (2018). UKParl: A Semantified and Topically Organized Corpus of Political Speeches. In D. Fišer & F. de Jong (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* Miyazaki, Japan, pp. 29–32.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, *45*(3), pp. 211–236.

Osborn, T. L. (2012). *How women represent women: Political parties, gender and representation in the state legislatures*. Oxford: Oxford University Press.

Pančur, A., Erjavec, T., Ojsteršek, M., Šorn, M., & Blaj Hribar, N. (2020). *Slovenian parliamentary corpus (1990-2018) siParl 2.0*. Slovenian language resource repository CLARIN.SI.

Poynton, C. (1989). *Language and gender: Making the difference*. Oxford: Oxford University Press.

Rix, K. (2014). ‘Whatever Passed in Parliament Ought to be Communicated to the Public’: Reporting the Proceedings of the Reformed C ommons, 1833–50. *Parliamentary History*, *33*(3), pp. 453–474.

Wängnerud, L. (1996). Intressen kontra stereotyper. Varför finns det kvinnliga och manliga politikområden i riksdagen? *Statsvetenskaplig Tidskrift*, *99*(2), pp. 129–149.

Wodak, R. (2007). Pragmatics and Critical Discourse Analysis: A cross-disciplinary inquiry. *Pragmatics & Cognition*, *15*, pp. 203–225.

Wolbrecht, C. (2002). *Female legislators and the women’s rights agenda: From feminine mystique to feminist era*. In Rosenthal, C. S. (ed.): Women transforming Congress. Norman: University of Oklahoma Press. 170–197.

1. The share is based on the official number of all female MPs listed in Kustec (2017) and then divided by the number of all MPs (i.e., 90). This share is slightly higher than reported in our analysis for which we strictly took only the corpus data throughout the tutorial. Since all of the elected MPs rarely serve the whole term, they are then replaced by other MPs. This is reflected in the corpus which includes all MPs, not just the ones listed at the time of the elections. According to corpus data, the total number of MPs per term is thus greater than 90. [↑](#footnote-ref-1)
2. Since this tutorial is offered both in English and Slovene, we gave the subcorpora in the concordancer bilingual names. However, for clarity and conciseness, we are only using the English part of the subcorpora names in this version of the tutorial. [↑](#footnote-ref-2)
3. Due to an error in corpus metadata attributed to one MP (specifically, to an MP named Geza whose gender was erroneously marked as female, most probably due to a combination of the ending –a, which in Slovene is typical of nouns of female grammatical gender, and low frequency of the name), the data for Term6, at the time of analysis, does not reflect the official number of female MPs which is according to Kustec (2017) 39 instead of 38 as found in the corpus. [↑](#footnote-ref-3)