Use Case 1
Gender in Parliamentary Discourse

Background

While more women than ever are being elected to parliaments around the world, equality is still a long way off, and current progress is far too slow. Most parliaments are still heavily male-dominated, and some have no women members of parliament at all. Even where women are present in greater numbers, glass ceilings often remain firmly in place. (Source: Women in Parliament by the Inter-Parliamentary Union)

CLARIN Resources and services used

- CLARIN Resource Families – parliamentary corpora
- the CLARIN.SI repository
- the noSketch Engine concordancer

Citizen scientist

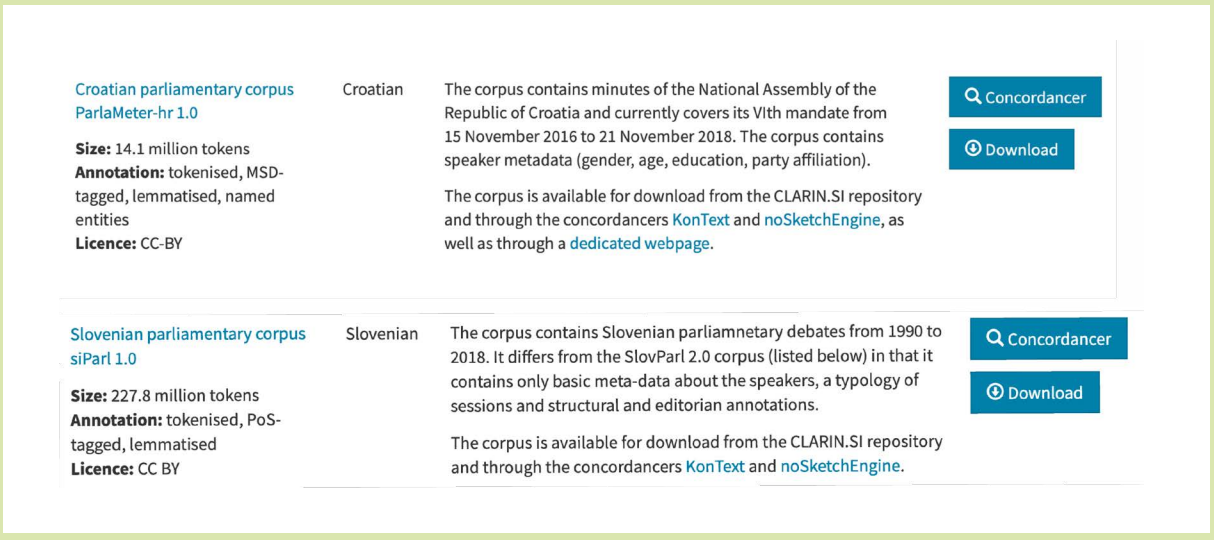
Do female speakers in the Slovenian and Croatian parliaments speak more or less than their male counterparts?

Student

Is the language of female parliamentary speakers similar in Slovenia and Croatia?

Step-by-step guide

1 Search the Parliamentary CLARIN Resource Family for relevant Slovenian and Croatian corpora. In this walkthrough, we'll use the Croatian and Slovenian ParlaMETER corpora, since they are roughly comparable in terms of time span, linguistic annotation and speaker metadata, but you can also use any of the other parliamentary corpora.



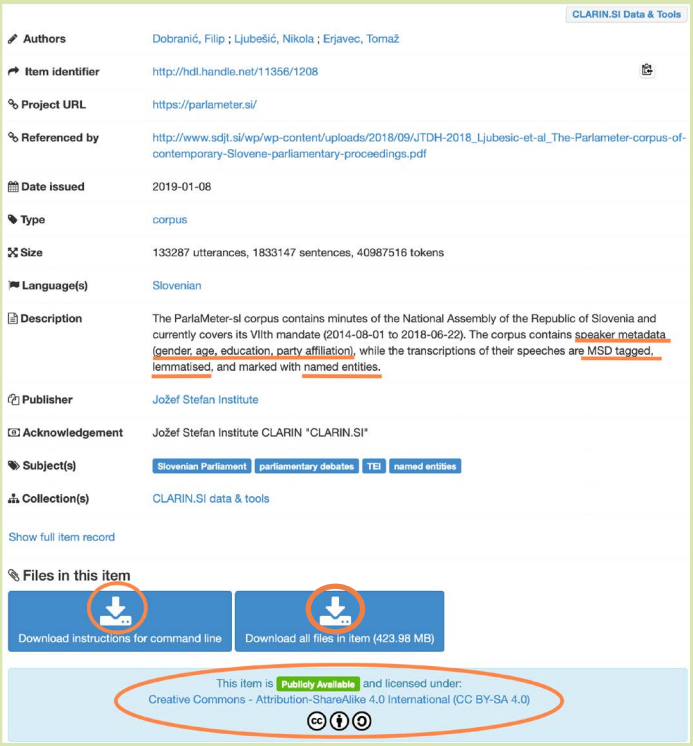
Croatian parliamentary corpus ParlaMeter-hr 1.0 **Size:** 14.1 million tokens **Annotation:** tokenised, MSD-tagged, lemmatised, named entities **Licence:** CC-BY	Croatian	The corpus contains minutes of the National Assembly of the Republic of Croatia and currently covers its VIth mandate from 15 November 2016 to 21 November 2018. The corpus contains speaker metadata (gender, age, education, party affiliation). The corpus is available for download from the CLARIN.SI repository and through the concordancers KonText and noSketchEngine, as well as through a dedicated webpage.	Concordancer Download
Slovenian parliamentary corpus siParl 1.0 **Size:** 227.8 million tokens **Annotation:** tokenised, PoS-tagged, lemmatised **Licence:** CC BY	Slovenian	The corpus contains Slovenian parliamnetary debates from 1990 to 2018. It differs from the SlovParl 2.0 corpus (listed below) in that it contains only basic meta-data about the speakers, a typology of sessions and structural and editorian annotations. The corpus is available for download from the CLARIN.SI repository and through the concordancers KonText and noSketchEngine.	Concordancer Download

CLARIN
Common Language Resources and Technology Infrastructure

SSHOC
social sciences & humanities open cloud

2 For both corpora, check their descriptions to see that:

- In terms of linguistic annotation, both corpora are **annotated for syntactic and morphological features** ("MSD-tagged"), **lemmatized** and **marked for named entities**.
- In terms of extra-linguistic annotation, both corpora are marked for **speaker metadata (gender, age, education, party affiliation).**
- The CC-BY licence shows that the corpus is publicly available, either for **download** or **on-line querying**.



3 Let's start by analysing the Slovenian corpus. First click on Slovenian parliamentary corpus ParlaMeter-sl 1.0 in the CLARIN Resource Families. This takes you to the record for this corpus in the CLARIN.SI repository:

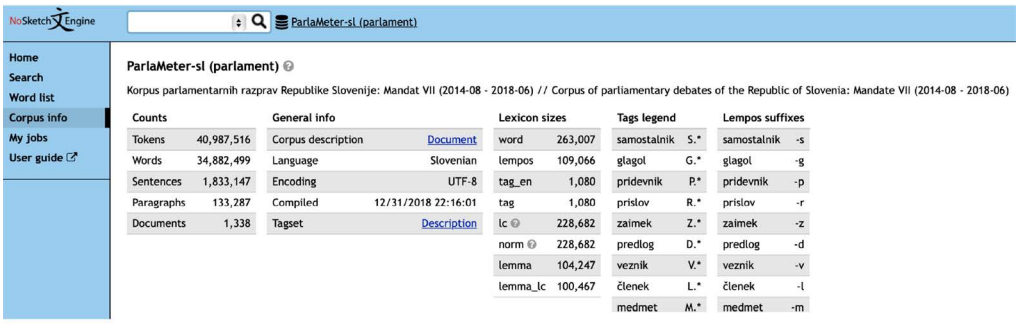
4 The CLARIN.SI repository shows how the corpus has to be cited to ensure proper author-ship attribution, and offers a **persistent identifier** for the resource – http://hdl.handle.net/11356/1208.

CLARIN
Common Language Resources and Technology Infrastructure

SSHOC
social sciences & humanities open cloud

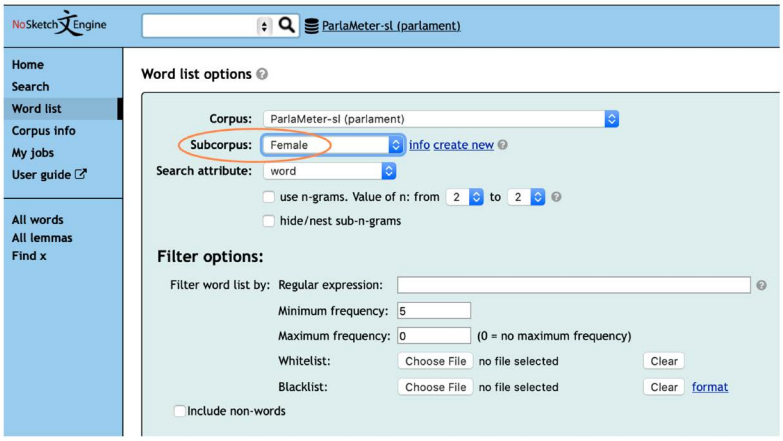
5 The corpus can be queried via two concordancers – **KonText** and **noSketch Engine**. Both offer very versatile search environments in which complex queries can be narrowed down on the basis of the **speaker metadata (age, party affiliation, etc).**

6 Let's query the corpus by using the **noSketch Engine**. In the repository, click on the downward arrow next to "noSketch" and then select search.



7 Let's recall our task: we're interested how parliamentary speakers are represented in the corpus in terms of gender. In other words, how many words of the total 34,882,499 are spoken by female parliament speakers and how many by male speakers?

8 We can figure this out by creating a **word list** and narrowing it down to the "Female" subcorpus.



9 After clicking on **Make word list**, we get the result for female speakers. We repeat the procedure for the "Male" subcorpus and see that the male speakers say 2.5 times more words than their female counterparts.

Repeat the procedure for the Croatian ParlaMeter corpus.

- What is the gender division in terms of words between male and female speakers in this corpus?
- Is the difference greater or smaller than that in the Slovenian corpus?

Additional Task

We can also construct word lists for individual word classes, such as nouns, verbs, adjectives, etc.

Which are the most frequent nouns used by the speakers in the Slovenian corpus?

a. Under search attribute, change from "word" to "tag_en". This specifies that you're searching for parts of speech rather than individual words.

b. In the filter option "Regular expression", write N.*. This specifies that you're searching for all nouns.

c. Under output options, select "lemma" under "Change output attributes". This ensures that all inflectional variants are all accounted for under a single base form of the word.

d. Click here to see the result for such a query.

Repeat the procedure for the Croatian ParlaMeter corpus. Are the results similar to the Slovenian corpus?

Research bite

In the Slovenian ParlaMeter corpus, the most frequent topics among the female speakers are ***health*** and ***labour, family*** and ***social affairs,*** which are followed by ***public administration and education, science and sport.***
Most of the 100 top-ranking keywords uttered by female speakers, on the other hand, could not be classified into a single topic because they were used either to achieve a ***stylistic effect***, were general words that were used in *multiple topics*, such as descriptive adjectives or legal terms, or ***ideological expressions,*** all of which indicate a more discursive, debating ***style*** of the male speakers, but could also stem from the fact that the leading roles in that term were predominantly held by male members of parliament (Source: Parlameter – a Corpus of Contemporary Slovene Parliamentary Proceedings by Darja Fišer, Nikola Ljubešić and Tomaž Erjavec).

Use Case 2

Creating a linguistically annotated corpus of 19th century English novels

Background

The digital humanities provide a new conception of the world of literature. Not only is this world larger – the sheer volume of the material we can access is unprecedented – but it is open to levels of analysis that could never be achieved by human brainpower alone. Hierarchies and themes fade into the background as patterns and networks emerge. These methods simultaneously divide texts into new categories and connect them to each other to form new wholes. (Source: When computers read: Literary analysis and digital technology by Sarah Jones)

CLARIN Resources and services used

- Virtual Language Observatory (VLO)
- Language Resource Switchboard
- WebLicht

Teacher

How can my students create an annotated corpus of 19th century English novels from scratch in an easy-to-use online environment?

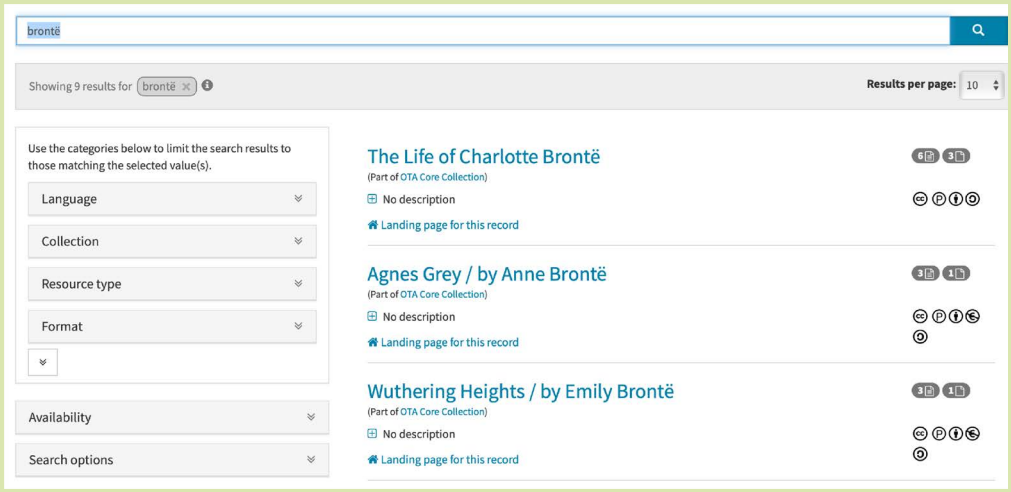
Researcher

Can you help me find resources and tools to research the stylistic differences between 19th century female and male novelists?

Step-by-step guide

The **Language Resource Switchboard (LRS)** aims at bridging the gap between resources (as identified in the VLO, Federal Content Search, and the CLARIN Virtual Collection) and tools that can process these resources in one way or another. For a given resource in question, it identifies all tools that can process the resource. It then sorts the tools in terms of the tasks they perform, and presents a task-oriented list to the user. Users can then select and invoke the tool of their choosing. (Source: Adapted from The Language Resource Switchboard by Claus Zinn)

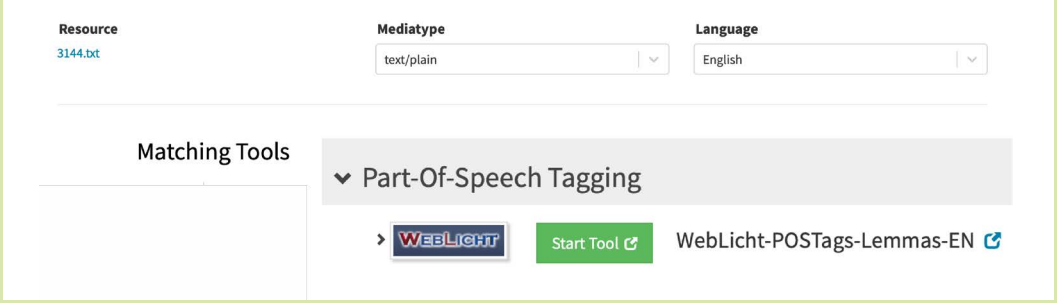
1 Search the VLO with the simple query Brontë.



This query gives you VLO records for 19th century English novels by the Brontë sisters, such as Wuthering Heights by Emily Brontë, The Tenant of Wildfell Hall by Anne Brontë, and Jane Eyre by Charlotte Brontë.

2 In each VLO record under the "Links" tab, we can find the complete novels in the form of .txt files. Each file can be processed through the Language Resource Switchboard by clicking on "…" next to the .txt file (in our case, 3144.txt).

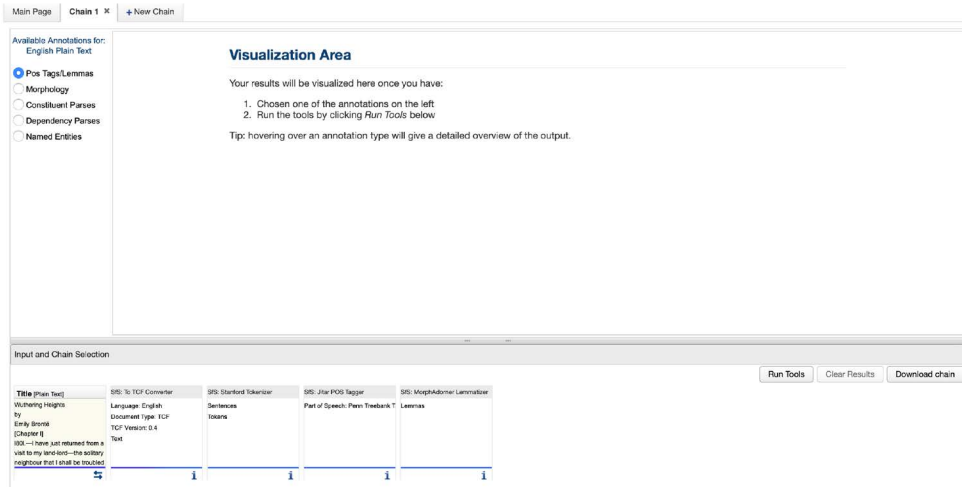
3 The first step in linguistic annotation typically involves **part-of-speech tagging**, with which each word in a corpus is assigned a part of speech, like *noun, verb*, and *adjective*. In the LRS, we see that part-of-speech tagging is performed by WebLicht.



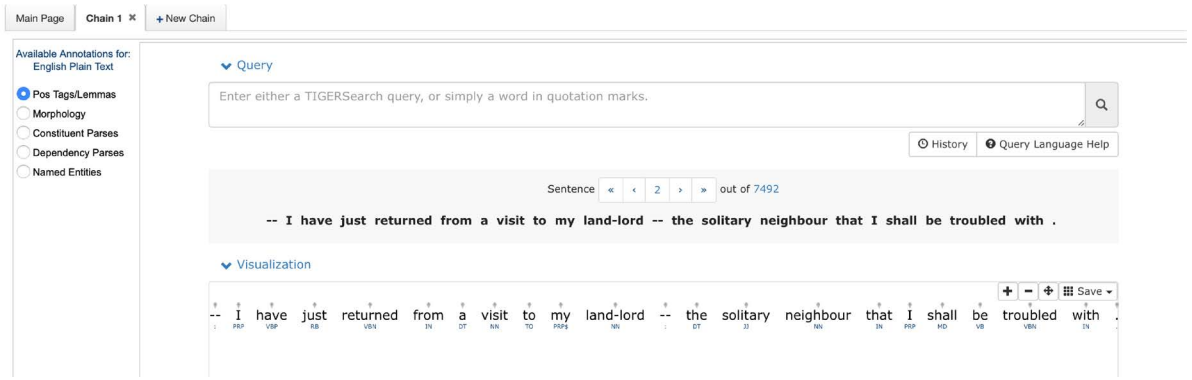
CLARIN
Common Language Resources and Technology Infrastructure

SSHOC
social sciences & humanities open cloud

4 In the WebLich application, we select *PoS tags/lemmas* under “Available Annotations for English Plain text”



5 After clicking on *Run Tools*, the entire *Wuthering Heights* novel becomes tagged for parts of speech.



The annotated novel can now be queried like a regular corpus either for simple words or by using the TIGERSearch corpus query language. To find all the adjectives in the newly tagged corpus, type [**pos** = /**J.***/] in the Query field. Make sure to enclose the query in square brackets. Try visualizing the results. Which are the most and least common adjectives in the novel? Hint: In the Statistics visualisation under “Add/remove columns”, try adding the values PoS and lemma.

WebLicht also allows you to create additional annotation chains (“New Chain”). Try to repeat the task above by also tagging The Tenant of Wildfell Hall for **parts of speech** as well as for **named entities**. By creating several annotation chains in this way, you are able to create a full-fledged linguistically annotated corpus, consisting of several novels which were originally in simple plain text.

Research bite

By analysing a corpus of novels by Charles Dickens, Mahlberg et al. (2019) have studied how fictional dialogue is used by the author to create a sense of realism and authenticity. The authors have shown that Dickens consistently writes dialogue characterised by linguistic features that fictional and real people share (e.g., question fragments, set expressions conveying politeness and vagueness), which contributes to a sense of naturalness to speech in fiction. By contrast, the range of frequent word combinations in fictional dialogue is more limited than that in spoken fiction, so it is possible that literature adds an iconic or heightened meaningful effect onto these forms. Particularly, shorter lexical combinations (e.g. *I mean, you know* and *I don't know*) are less frequent in fiction. (Source: Speech-bundles in the 19th-century English novel by Michaela Malmberg, Viola Wiegand, Peter Stockwell, and Anthony Hennessey)