CLASSLA and ReLDI

Regional variation in gender marking

Worksheet for a hands-on tutorial on extracting data from corpora

Gender marking is a topic that sets the study of language in a broader contexts of studying society and its complex relations. In South Slavic languages, for instance, gender appears in two forms:

- Grammatical: e.g. *stolica* (fem. 'chair') vs. *sto* (masc. 'table'). The real world objects denoted by these nouns do not have any gender, but the words themselves do show grammatical gender. We can tell the grammatical gender by the form of the words (ending *-a* is typical for the feminine gender) and by the ending of their modifiers (*velik**a** stolica* 'big chair' vs. *velik**i** sto* 'big table').
- Natural: e.g. *direktorica* (fem. 'director') vs. *direktor* (masc. *director*). In this case, the form of the noun is determined the gender of the person whose occupation is denoted.

Recently, there has been a lot of discussion on the use feminine forms to denote female occupations, to the point that this has become regulated by the law in some countries. On the other hand, the current biases in the society can be studied by looking at how people speak or write. Are there really many more male (*direktor*) than female (*direktorica*) directors in the public communication. How many more? Do we see stronger biases in more prestigious occupations? How about the regional variation; are biases stronger in some regions than others? Some answers to such questions can be found in large language corpora.

The goal of this tutorial is to help researchers make the first steps in collecting and analysing data from language corpora. It is intended for researchers in various fields who would like to be able to consult large language corpora for their studies on different topics. The tutorial should serve as an example of how corpus searches are performed, but also, more importantly, on how these searches should be integrated in the research process: how to go from a research question to a corpus query, where to find and how to select relevant corpora, what counts to extract and how to interpret them. We take gender marking in South Slavic languages and the questions listed above as examples of research questions that can be approached with a quantitative corpus study. We hope that these examples can inspire and facilitate the use of large language corpora for future studies of the relation between language and society.

Part one: Selecting sources

---

1. Find relevant corpora using CLARIN

The best way to find feely available and well encoded corpora is via the CLARIN infrastructure (https://www.clarin.eu). This is a wide European network of data repositories dedicated to language resources.

If you are interested in South Slavic languages, as we are in this tutorial, bookmark the CLARIN.SI starting page: http://www.clarin.si/info/about/ (http://www.clarin.si/info/about/). This is a CLARIN centre in Slovenia specialised in (but not limited to) South Slavic languages.

Our goal in this tutorial is to cover closely related languages of former Yugoslavia. Assuming that the regions inside this territory are defined as former republics and current states, each with its own web domain, list all the corpora on CLARIN.SI for each region:

1. Bosnia

2. Croatia

3. Montenegro

4. Serbia

5. Slovenia

2. Select large corpora of the same type: one for each region

From the lists above, chose one corpus to search in the following exercises. Try to find the same kind of corpus for each region. After deciding what corpora to use, fill in the following table with links to the corresponding corpus search forms:

	Corpus name	NoSketchEngine concordancer	KonText concordancer
Bosnia			
Croatia			
Montenegro			
Serbia			
Slovenia			

3. Observations

Note down here any observations about completing Step 1 and Step 2. How did you find the list of resources? How did you find the concordancers? What were the obstacles? What would you do differently next time?

Part two: Occupations and social prestige

---

4. Feminine vs. masculine nouns describing occupations with higher social status.

4.1 Find the following raw counts (number of occurrences) for feminine and masculine nouns for each set of nouns in each region:

	M (direktor)	F (direktorka)	F (direktorica)
Bosnia			
Croatia			
Montenegro			
Serbia			
Slovenia			

4.2 Calculate the ratio of feminine to masculine nouns for each set of nouns in each region:

	M (direktor)	F (direktorka)	F (direktorica)
Bosnia			
Croatia			

	M (direktor)	F (direktorka)	F (direktorica)
Montenegro			
Serbia			
Slovenia			

4.3 Observations

Note down here any observations about completing Step 4.1 and Step 4.2. What trends do we see in raw counts? What trends do we see in the ratios? Describe the differences between regions.

5. Feminine vs. masculine nouns describing occupations with lower social status.

5.1 Find the following raw counts (number of occurrences) for feminine and masculine nouns (Engl. *teacher*) for each set of nouns in each region:

- `nastavnica - nastavnik` (bsWaC, hrWaC, meWaC, srWaC)
- `učiteljica - učitelj` (slWaC)

	F	M	Total
Bosnia			
Croatia			
Montenegro			
Serbia			
Slovenia			

5.2 Calculate the ratio of feminine to masculine nouns for each set of nouns in each region:

	F	M	Total
Bosnia			
Croatia			
Montenegro			
Serbia			
Slovenia			

5.3 Observations

Note down here any observations about completing Step 5.1 and Step 5.2. What trends do we see in raw counts? What trends do we see in the ratios? Describe the differences between regions.

Describe the differences between the trends in Step 4 and those in Step 5. Were there any surprises? What would be your next search?

Part three: Semantic and social roles

---

6. Find the ratio of feminine to masculine subjects in each region for the following three verbs: misliti (think), os(j)ećati (feel), izjaviti (state, communicate to the media).

	misliti	os(j)ećati čutiti	izjaviti	

	M (misliti)	F (misliti)	M (os(j)ećati/čutiti)	F (os(j)ećati/čutiti)	M (izjaviti)	F (izjaviti)
Bosnia						
Croatia						
Montenegro						
Serbia						
Slovenia						

Tips:

- Think how to represent the gender of subject using morphosyntactic definitions (MSD) of verbs
- You don't need all instances of all subjects to calculate this ratio
- Think what raw counts you need for this calculation

7. Describe the trends in Section 6.:

7.1 Are there any regional patterns?

7.2 Are there any other patterns?

7.3 What do they mean?

7.4 How do they compare to the findings in Section 4 and Section 5?

*8. Use the "Frequency" button in the NoSketchEngine menu on the left side of the interface to visualise the ratios in 6 as horizontal bar plots. Capture your plots and include them here:

* increased difficulty

Tip: You need a single query for both forms (masculine and feminine)

8.1 'misliti'

8.2 'os(j)ećati'

8.2 'izjaviti'

Part four: Summary and conclusions

- What do corpora tell us about gender bias in the society?

- What have we learned about regional variation?

- Are there any surprising findings? If yes, why are they surprising?