Daria Lazić, PhD Student
Institute of Croatian Language and Linguistics
Zagreb, Croatia

Report on Elexis Transnational Research Visit Grant at Det Danske Sprog- og Litteraturselskab and University of Copenhagen

(Copenhagen, Denmark, June 16 – July 5, 2019)

Project title:

Nordic E-dictionaries in Comparison to the *Croatian Web Dictionary – Mrežnik*

Introduction

I applied for a visit to Det Danske Sprog- og Litteraturselskab and the University of Copenhagen with two main goals in mind. Firstly, since I am working on a project within which a new Croatian web dictionary is being compiled, I wanted to visit an institution where a similar project is being conducted in order to exchange ideas and discuss possible common issues. In that respect, Det Danske Sprog- og Litteraturselskab (DSL) was interesting because of *Den Danske Ordbog*, a comprehensive corpus-based dictionary of modern Danish. Secondly, I sought to conduct an initial research for my PhD thesis, in which I will be comparing the work on the Croatian web dictionary with similar foreign dictionaries, in particular dictionaries of Nordic languages. Among other things, I have been interested in the way sensitive content, such as gender, religion, nationality and other, is represented in corpora and dealt with in dictionaries.

The visit has been useful and informative in all the aspects mentioned above. I was introduced to DSL's staff and the projects they are working on and later I had meetings with several editors of *Den Danske Ordbog*. Besides that, everyone was very helpful and eager to answer my questions. I was provided with research material on the topics I am interested in and had an opportunity to explore DSL's tools and resources on my own.

Furthermore, I visited the Centre for Language Technology at the University of Copenhagen, met its researchers and was introduced to their projects, tools and resources. At the end of the visit, I presented the Croatian dictionary project I am working on to the editors of *Den Danske Ordbog*.

In addition, it proved to be a good time to visit due to several events relevant for my work that took place at DSL during my stay there, such as a visit of the lexicographers from the Centre for Digital Lexicography of the German Language and a guest lecture on ethics in lexicography.

Below I will discuss some of the activities mentioned above in more detail.

Scheduled meetings with editors of *Den Danske Ordbog*

During my visit, I met with several of the editors of *Den Danske Ordbog* and discussed diverse topics with them. Some of these are the following:

- introduction to DSL and an overview of their **projects and resources**, in particular *Den Danske Ordbog* (*DDO*; a dictionary of modern Danish, an ongoing project nowadays published online at ordnet.dk/ddo) and *Den Danske Begrebsordbog* (a Danish thesaurus)
- introduction to the **tools** used by the lexicographers at DSL:
    - corpus query system CoREST
    - Word2Dict tool for lemma selection
    - dictionary writing system iLex

    → I was provided with access to the tools as well as the resources which are being edited in iLex, primarily *DDO*, and I had an opportunity to study them on my own. A feature that I found especially useful was the search function in iLex which enables an easy retrieving of information from the dictionary, such as all entries containing a certain usage label, word or semantic field.
- work on *DDO*:
    - **XML structure of the articles** in iLex and the information they contain

→ an interesting feature is for example that the word senses are equipped with genus proximum and an id-number that they share with other lexical resources developed at DSL

- **lemma selection**: candidates for lemmas that could be included in the dictionary can be found in the CoREST corpus tool:
    - frequency word lists generated from the corpora: it is indicated whether a certain word exists in the dictionary, whether someone is already working on the entry or it is free for a lexicographer to compile it
    - suggestions from users together with their comments
    - Word2Dict – a tool that presents semantically related words and indicates whether each of them exists as a lemma in *DDO*; for the lemmas that have already been included in the dictionary the definitions are shown and in that way the tool assists the lexicographer both in selecting new lemmas and writing consistent definitions of lemmas that are semantically related
- **variants in the dictionary**: the dictionary is corpus-based, which means that orthographical, morphological and other variants sometimes appear in the corpus material; both corpus frequencies and conventions proposed by the Danish Language Council are taken into account and the differences are brought up in the dictionary when relevant; since the dictionary examples are taken from the corpus, variants such as different spelling can appear in them
- **revising the dictionary**: along with expanding the dictionary with new lemmas, the existing ones are revised; some of the elements that are revised are spelling (for example in loanwords), definitions and examples (often pointed out by users as problematic for some reason), and lemma inventory (for example, some words that denote phenomena in society or technology can be outdated); currently, an effort is being made to revise controversial words and expressions such as derogatory words and words related to certain social groups or sensitive topics (nationality, religion, gender, age, disabilities, physical features, etc.)

→ practical solutions in *DDO*: an article can be marked for revision at a later point (for example words in the field of technology); genus proximum allows selection of a certain semantic group for revision; example hierarchy – when a new example is added, the older ones can be downgraded or removed

- **stereotypes and potentially offensive content** in the corpus and in the dictionary: problems regarding the presentation of sensitive content (which is commonly related to minority groups in the society) can appear when selecting lemmas for the dictionary, defining their meanings and usage and exemplifying them; part of the problem is that the corpora dictionaries are based on are not always free from stereotypes and offensive language use nor they include texts that specifically regard minority groups; offensive words and expressions are especially problematic because their inclusion in the dictionary is often perceived by the public as an approval of their use

    → I was introduced to the latest discussions about the representation of minority groups in *Den Danske Ordbog*, to reactions from dictionary users and changes they have resulted in

    - among other things, the use of usage labels and explanations boxes has been discussed

    → I was provided with research articles on the topic as well as internal lists of problematic words compiled at DSL, which will be useful for me when studying the way such content is presented in Nordic and Croatian dictionaries and dealing with it in my own work

- **reactions from the dictionary users**: users can both leave comments under a certain article or contribute by suggesting words and expressions that should be added to the dictionary (*Spordhund* function); a list of recent words suggested by users that have been added to the dictionary is published and in that way users are encouraged to participate
- **other resources** developed at DSL:
    - *Den Danske Begrebsordbog*: I have been introduced to the Danish thesaurus, its structure and content, as well as the work on its extension and the logic behind its integration into *Den Danske Ordbog* in the form of

    related words, *ord i nærheden*; I also had an opportunity to browse through the XML document and the printed version of the dictionary
    - *Svensk-Dansk Ordbog* (*Swedish-English Dictionary*) – printed and online version

Visit to the Centre for Language Technology (University of Copenhagen)

During my stay in Copenhagen, I also visited the Centre for Language Technology (Center for Sprogteknologi) at the University of Copenhagen, where I met with the researches who introduced me to their main projects, among which are the following:

- *DanNet* – the Danish WordNet that has been compiled based on the senses in *Den Danske Ordbog* in collaboration with DSL
    - linking to other resources: the researchers explained and showed me the process of linking *DanNet* to *Princeton WordNet*, a project they are currently working on, as well as WordTies, a multilingual WordNet browser for Nordic and Baltic languages
- *The Danish FrameNet* – based on the Berkeley FrameNet model

Furthermore, the researchers at the Centre for Language Technology offered to train two of their tools – lemmatizer and POS-tagger – for Croatian, and I plan on comparing them with the tools already available for Croatian shortly.

Other activities

Among other activities that took place during my visit, it was very interesting to take part in the visit of lexicographers from the Centre for Digital Lexicography of the German Language (https://www.zentrum-lexikographie.de) and follow a seminar where revising and updating the dictionary and communication with the general public were the two main discussion topics. It was a unique opportunity to get an insight into another lexicographic project and follow a discussion on lexicographic problems and solutions.

Another lucky circumstance has been that I could attend a visiting lecture by Dr. René Rosfort from the University of Copenhagen on ethics in lexicographic work, a topic that is both one of my research interests and has proven to be actively discussed among the editors of the Danish dictionary, often encouraged by its users. The lecture sought to explain the problems behind the description of sensitive words and many interesting examples of such content were mentioned.

Finally, at the end of my visit I presented the *Croatian Web Dictionary – Mrežnik* project I am working on to the editors of *Den Danske Ordbog* and brought up a couple of problems that we are facing in our work. The presentation was followed by an interesting discussion which provided useful feedback for my future work.

Conclusion

Looking back on my research visit, I can say that it succeeded beyond my expectations and I believe that the experience and contacts I have gained from it will be extremely valuable for my future work. Over the course of three weeks spent in Denmark, I met researchers from both Det Danske Sprog- og Litteraturselskab and the Centre for Language Technology at the University of Copenhagen and I was introduced to their work. I gained an overview over Danish lexical resources and an in-depth insight into *Den Danske Ordbog*, a resource most relevant for me. Through informative and inspirational conversations with the editors of the Danish dictionary, I became acquainted with the current discussions and trends in Danish and Nordic lexicography, and I was provided with ideas for my own work.

I would like to use this occasion to thank my hosts – all the DSL staff – for their hospitality and for making me feel at home during my stay in Denmark. Furthermore, I would like to thank all the editors and researchers at both Det Danske Sprog- og Litteraturselskab and the University of Copenhagen for eagerly answering all of my questions and providing me with additional material on the topics I found interesting. I would, in particular, like to express my thanks to Dr. Sanni Nimb, Senior Editor at DSL, for assisting me prior to and during my visit as well as planning my activities, and to Sussi Olsen from the University of Copenhagen for organizing my visit to the Centre for Language Technology and for administrative assistance. Finally, I am grateful to the Elexis project for making my research visit possible.