Eglantina Gishti, Professor of Linguistics
University of Tirana
Tirana, Albania

Report on Elexis Transnational Research Visit Grant at Det Danske Sprog- og Litteraturselskab and at the Centre for Language Technology, Department of Nordic Studies and Linguistics, University of Copenhagen (Copenhagen, Denmark, April 4 – April 8, 2022)

Project title:

A Corpus-based method for Extraction of Polylexical Units

(in French and Albanian languages)

Introduction

I applied for a visit to Det Danske Sprog- og Litteraturselskab and to the University of Copenhagen to discuss various aspects of the work related to dictionaries and corpora and to know the functions of all the corpus they created and their specificities. Since my research concerns lexicography and NLP and being aware of the lack of lexicography work and products in Albania, I wanted to visit an institution where a lexicographic project is being conducted (new lexicographical methodology) in order to exchange ideas and discuss different issues concerning the web dictionary. In that point of view, at Det Danske Sprog- og Litteraturselskab (DSL) I was introduced with the project ordnet.dk, a comprehensive corpus-based dictionary of modern Danish and other important projects ongoing in Det Danske Sprog- og Litteraturselskab (DSL).

Furthermore, I visited the Centre for Language Technology at the University of Copenhagen, met the researchers, and was introduced to their projects, tools and resources. My initial idea was to conduct research, in which we will create parallel data on foreign languages and Albanian language that will allow us to do research and to establish a rich database for the language technology. The visit at the University of Copenhagen made me aware that it will be more convenient to start with a monolingual corpus because this will enable us to have more complete data in the respective languages. Once it is done, the work on parallel corpora will be easier. Based on the experience in both institutions, I realized the importance of the lexicon represented in corpora and dealt with in dictionaries. In my project interest, this is important as we work as well

with the Extraction of Polylexical Units that are difficult to be presented in a dictionary. The reference to the Corpora is a must in that case.
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 of the Centre for Language Technology at the University of Copenhagen. Besides that, everyone was very helpful and eager to answer my questions.

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

4 – 5 April 2022: visit to the Centre for Language Technology (University of Copenhagen)

During my stay in Copenhagen, I started the visit at the Centre for Language Technology (Center for Sprogteknologi) at the University of Copenhagen where I was welcomed by Prof. Bolette Sandford Pedersen and Sussi Olsen. They presented me to the other staff members and the researchers who introduced me to their main projects, among which are the following:

- *CLARIN-DK* as an infrastructure where researchers can deposit, share, and download language-based material. i.e., texts, transcriptions, lexicons, word lists, audio, and video files. CLARIN-DK also comprises interactive language tools.
- *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.
- *The Danish FrameNet* – based on the Berkeley FrameNet model
- *ParlaMint* - a project which contributes to the creation of comparable and uniformly annotated multilingual corpora of parliamentary sessions. This was interesting w.r.t. the idea of parallel corpora idea.
- ONP: Dictionary of Old Norse Prose - http://onp.ku.dk/
- Wordnet tool
- Web scraping methods

Furthermore, the researcher at the Centre for Language Technology offered to start **training CLARIN tools f**or Albanian.

6-8 April 2022 - Scheduled meetings at Society for Danish Language and Literature

During my visit, I spent the rest of my mobility at the Society for Danish Language and Literature, and I was welcomed there by Sanni Nimb. 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) and *Den Danske Begrebsordbog* (a Danish thesaurus)

- I was also introduced also to Ordbog over det danske Sprog (ODS; a dictionary of older Danish published online at ordnet.dk/ods), because of my interest in historical texts – not only lexicography and modern corpora.
- Concerning the building of monolingual corpora: I discussed with Ida Flörke the issues of collecting text material from the publishing houses and she showed how the juridical contract is formulated. She also provided me with a version in English.
- introduction to the **tools** used by the lexicographers at DSL:
    - dictionary writing system iLex
- 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 (**tool for linguistic studies** in very large text collections).
        - frequency word lists generated from the corpora
        - 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 define.
- **other resources** developed at DSL:

I also met the director of DSL, Dr. Karen Skovgaard-Petersen and Finn Gredal Jensen. They showed me some of their work. One of these works is the online Holberg edition: http://holbergsskrifter.dk. And we exchanged experiences on our current work.

Dr. Marita Akhøj Nielsen told me about Leonora Christina's French autobiography and Music and language in the Danish hymn singing of the Reformation period.

In addition

The last day of my mobility, we had a visit at the Royal Library, guided by Dr. Anders Toftgaard. This visit was possible because DSL collaborates closely with the library. We profited to see among different collections: the Royal Library's Manuscript Collection contains manuscripts which range from the early Middle Ages to the present. We were introduced to a lot of information concerning the library, its resources, and its history.

We also discussed the lexicographical situation of the Albanian language, my work on corpora and dictionaries. We exchanged different ideas about how to improve the lexicographical work, about the importance of the books and manuscript's digitization, etc.

Conclusion

The visit proved to be above my expectations. From historical data and dictionaries to different tools, the projects I was introduced revealed to be a true inspiration and to rethink the methodology I had conceived for my project. I believe that the experience and contacts I have gained from it will be very valuable for my work. 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 the tools and Danish lexical resources. Through informative and inspirational conversations with the editors of the Danish dictionary and the researchers, I was provided with ideas for my own work.

I would like to use this occasion to thank my hosts – Centre for Language Technology (University of Copenhagen) and 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 researchers and editors at both institutions for explaining to me all their work and answering all my questions. I would, in particular, like to express my thanks to Dr. Sanni Nimb, Prof. Bolette Sandford Pedersen and to Sussi Olsen, for assisting me prior to and during my visit as well as planning my activities and for administrative assistance. Finally, thank you to the Elexis project for making this research mobility possible.

11/4/2022