Annika Simonsen, linguist
Talutøkni, the Faroe Islands

Report on Elexis Transnational Research Visit Grant at Det Danske Sprog- og Litteraturselskab (DSL, Denmark) & University of Copenhagen (UCPH, Denmark)

(Copenhagen, Denmark, November 16th – 27th, 2021)

I was given the opportunity to visit Det Danske Sprog- og Litteraturselskab (DSL, Denmark) and the Centre for Language Technology (CST, Denmark) at the University of Copenhagen (UCPH, Denmark), who together with Dansk Sprognævn (DSN, Denmark) form one of the most important research centres for NLP in the North, specialising in language technology for the West Nordic languages. During my visit, I met with several experts, who shared their valuable knowledge with me. Aside from getting assistance with my own work, I made several contacts at both DSL and CST, and opened up the possibility for future collaboration.

My project, Project Ravnur, is a Faroese speech recognition project that creates versatile language materials for a broad range of Faroese LT (a so-called BLARK, or Basic Language Resource Kit). My current role as linguist in our project is to oversee our transcription assistants, but I am also creating a wide-coverage dictionary including phonetic information for all word forms and Part of Speech (PoS) tags. My research visit was originally planned to be in fall of 2020. My task at the time was to prepare a PAROLE tag definition for Faroese and apply the tagset in the dictionary and transcription corpora. I applied to visit DSL and CST at UCPH because they are the Danish partners in the PAROLE-project; this made them my ideal hosts.
My visit was pushed back a year and by that time I had already completed my PAROLE tag definition and had moved onto other things. However, I still had two things I wanted to learn from DSL and CST:

1) Det Centrale Ordregister (COR)

I am currently working on a GUID-style index system called OTAL (O-number) to use for all linguistic data in Project Ravnur. The index system is developed for implementing the system into the new Faroese orthographic dictionary, when it has been completed. The index system takes its main inspiration from the ongoing Danish project, COR (the Central Danish Word Register). COR is a collaboration between DSN, DSL and CST. It is an ongoing project, which means I would get to witness the work in progress. Furthermore, I could get assistance with the Faroese OTAL.

2) General insight and contacts

My second goal that I had in mind was to get general insight into as many relevant projects at DSL and CST as possible. Both institutions are home to some of the leading researchers in the North. Being able to shadow their work would be a valuable opportunity, and I would strive to create contacts for Project Ravnur.

Det Danske Sprog-og Litteraturselskab (DSL)
At DSL I was introduced to several dictionary editors and their work. I was given the opportunity to explore DSL's tools and resources on my own. My first meeting was with a senior editor, who introduced me to DSL's corpus tool, CoREST, and the PAROLE tags that are used in it. I learned how DSL obtains their text for the corpus and how their metadata is stored. I am interested in corpora and text collection, because Faroese lacks a big text corpus and Project Ravnur is currently in the process of creating one for our language model. It was interesting to see how CoREST has been created and to learn about the potential challenges that come with maintaining a big text corpus.
Other relevant resources that I was given more insight into was Den Danske Ordbog or DDO (an online dictionary of Modern Danish) and its structure, as well as the Den Danske Begrebsordbog (the Danish thesaurus), which categorises words according to their relatedness - something that is integrated in DDO as well as in other projects.

Not only did I learn about DSL's own projects during my visit, but I was also given the opportunity to show them my own work and ask for advice. It is not often that I get to ask for advice from someone who has experience both with lexicography and language technology like the editors at DSL do. The feedback I received on our ongoing OTAL project was very helpful. Furthermore, now I have made contacts who are familiar with my project and whom I can reach out to next time I have any questions.

The Centre for Language Technology (CST) at the University of Copenhagen (UCPH)
At CST I was introduced to the research staff and their work. It would not be possible for me to mention every single thing I learned at CST, so I will only include the highlights. Among the many relevant resources I learned about was DanNet, a wordnet for Danish. DanNet shows how concepts relate to other concepts, e.g. that cake (*kage*) is used for eating (*spise*) and is made of sugar (*sukker*) and flour (*mel*).

Some researchers at CST even showed an interest in developing tools for Faroese. I provided them with the necessary materials and they trained a lemmatizer (CSTlemma) on Faroese. As of writing, we are testing several tools for Faroese and we are working together to improve them. This was an unexpected, but welcomed opportunity for me, and I am delighted to have Faroese LT tools in the making. Furthermore, a research assistant at CST generously donated her time to program scripts for me to use to collect Faroese news text from the internet. These scripts are making it possible for us in Project Ravnur to proficiently collect text for our background text corpus, which we will use for our language model.

Other activites
I was lucky to part-take in several extra activities during my research visit. At the very beginning of my visit I got to present a poster about Project Ravnur and OTAL at the Language Technology conference (Sprogteknologisk Konference) at CST. Later I attended professorial inauguration lectures in Danish language at UCPH, as well as a virtual European Language Resource Coordination (ELRC) conference. I also managed to take a day-trip to Bogense to visit Dansk Sprognævn (DSN). And finally, on one of my last days, I was given a tour of historical Faroese documents at the Arnamagnæan Manuscript Collection, which were being restored.

Conclusion
As someone working with an extremely low resource language such as Faroese, it was helpful to visit institutions who also work with a low resource language such as Danish, because I was able to learn from experienced people who have faced similar challenges as myself. There is no established Language Technology (LT) field in the Faroe Islands as of yet, but it is easy to recognize the need for one. With the arrival of the digital age, Faroese could be facing digital extinction if it cannot keep up. Developing good LT resources for Faroese will therefore play a crucial role for the long-term survival of Faroese. The things I have learned at DSL and CST are going to guide me when I create resources for Faroese, because I have seen how the Danish resources have been made and I have seen the challenges they have faced and how they solved their problems. On top of having learned a lot, I now know several familiar faces at DSL and CST, who I can contact next time I need guidance. I am planning to stay in touch.

I would like to thank Bolette Sandford Pedersen and Sussi Olsen at CST and Sanni Nimb at DSL for their wonderful efforts while being my hosts, and also every expert at DSL and CST who shared their knowledge and experience with me, especially Thomas Troelsgård at DSN for helping me with OTAL; Bart Jongejan and Nathalie Hau Sørensen at CST for creating LT tools for Faroese with me; and Peter Juel Henrichsen at DSN for encouraging me to apply for the ELEXIS grant.