**ELEXIS Transnational Research Visit Grant**

**Final report**

**Grant holder:** Tanara Zingano Kuhn, PhD

**Affiliation**: Centre for the Study of General and Applied Linguistics at the University of Coimbra, Portugal

**Host Institution**: Institut Jožef Stefan (JSI, Slovenia)

**Host**: Iztok Kosem, PhD

**Period:** 25/6/2019 to 10/7/2019

**Project Title**: Improving a procedure for automatic extraction of data and import into DWS

## 1. Purpose of the project

One of the objectives of ELEXIS is to develop tools which can be used by all European institutions working with lexicography in order to promote common standards. One of these tools is the Lexonomy dictionary writing and online dictionary publishing system, which interacts with Sketch Engine for automatic access to corpus data, thus streamlining the process of entry writing and editing.

While in my PhD (Kuhn, 2017) I developed a procedure for automatic extraction of data from a corpus and import into a dictionary writing system (DWS) in order to create a design of a dictionary of Portuguese for university students, my current research focuses on making a prototype for this dictionary, however, this time using Lexonomy. That means that, instead of having to go through a laborious and lengthy two-step process – first extraction of data from the corpus, then import into DWS -, now all this process takes place directly in Lexonomy. This is because, as informed on the ELEXIS website, "Sketch Engine can push lexicographic data into Lexonomy to create automatically-generated dictionary drafts and Lexonomy can pull data from Sketch Engine's corpora during the entry editing process."

Thus, it was my purpose during this ELEXIS transnational research visit to contribute to the development of Lexonomy by working on entry modelling. As part of a multilingual team of lexicographers, I was responsible for bringing up some characteristics of the Portuguese language. More specifically, I focused on special lexicographic needs that derive from a

perspective of Portuguese as a pluricentric language. My research questions were: how can Portuguese as a pluricentric language be catered for in a model entry? What is necessary to link VOC (a very rich lexical database of Portuguese) with my dictionary prototype on Lexonomy?

One of the goals of this team of lexicographers is to propose a Lexonomy model entry that is as comprehensive as possible so that dictionary makers of any languages can use Lexonomy for their projects. In my research visit, I worked towards helping them achieve that.

## 2. Description of work carried out during the Research Visit

Firstly, I was introduced to the Lexonomy project in more detail so that I could fully understand the extent to which I could contribute, as well as tune in with their needs. I learned about a very innovative lexicographic project that is currently being developed for Slovene in which Lexonomy is used as DWS. In this project, the functionalities for interaction with Sketch Engine are being successfully adopted, indicating that the tool can be used for the development of other dictionary projects already. In addition to this project, Iztok Kosem gave me an enlightening introduction to a series of other pioneering Slovene lexicography projects that involve, among other state-of-the-art techniques, publication of automatic extracted data in dictionary entries, development of games-with-a-purpose word games and use of crowdsourcing for supporting entry writing. I was thrilled by so many ground-breaking projects and already started conversations with Iztok in order to apply such techniques to the lexicographic projects of Portuguese with which I am involved, such as the development of my dictionary prototype.

One of the paramount conditions for working on entry modelling is to have a common understanding of what each entry element consists and to establish standard terms to refer to them. Thus, I was given access to a shared document regarding literature review on which the above-mentioned group of multilingual lexicographers had been working, which gave me the opportunity to learn from this rich material as well as to contribute to it by doing a series of further specialised readings.

Iztok Kosem and I then had several meetings to discuss what elements should be comprised by the Lexonomy model entry. These were unique moments that gave rise to enlightening theoretical discussions on language, language science and metalexicography. One of our topics of discussion concerned the fact that an entry for a dictionary of Portuguese as a pluricentric language must cater for the fact that there is variation in Portuguese as a result of where it is spoken. I introduced VOC – Vocabulário Ortográfico Comum da Língua Portuguesa (The common spelling dictionary of the Portuguese language) http://voc.cplp.org/ to Iztok Kosem

so that he could help me find a solution to link the VOC database to my dictionary prototype in Lexonomy.

## 3. Description of the main results obtained

At the end of my visit, Iztok and I wrote a model entry in XML that comprised a variety of elements structured in a well-thought-out manner and sent it to the core team of computational linguists working on the development of Lexonomy.

The special case of linking VOC with my dictionary in Lexonomy was carefully debated, however, we could not come up with a final result given lack of time. It is worth mentioning that Iztok showed great interest in this topic, so we will certainly continue research on this issue in a future work.

Even though my contribution is only a small part of a much larger project, I am glad to know that the actual needs concerning the Portuguese language will be considered when creating a model entry for Lexonomy. As a result, I will be able to carry out my project of a dictionary of Portuguese for university students using the most advanced techniques for dictionary making. Moreover, any other lexicographer working on dictionaries of Portuguese will benefit from this ground-breaking tool.

## 5. Concluding remarks

This Transnational Research Visit Grant has been highly instructive and motivational. I have had the opportunity to learn about pioneering lexicographic projects for Slovene, which not only contributed to knowledge growth, but also served as an incentive for me to share this information with colleagues at my host institution. I am highly motivated to further learn about those techniques in order to work on adjustments for implementation of projects related to Portuguese.

It should be highlighted that those meetings with Iztok Kosem took place in different locations, namely, Institut Jožef Stefan, Trojína Research Centre and University of Ljubljana. Such variety allowed me to visit different laboratories and research institutes and get to know researchers working on a number of lexicographic projects. There is no doubt that this was a great opportunity for expanding my academic network.