**Host Report of Short Term Scientific Mission**

**COST STSM Reference Number:** COST-STSM-IS1305- 210216-071459

**Period:** 21/02/2016 to 30/04/2016

**Duration**: 50 working days.

**COST Action:** IS1305

**STSM type:** Regular (from Portugal to Slovenia)

**STSM Title**: In-depth Introduction to Automatic Knowledge Acquisition for Lexicography and DWS

**Guest/STSM applicant**: Tanara Zingano Kuhn – Capes Scholarship holder nº 0973/13-0 /University of Lisbon/CELGA-ILTEC

**Host**: Iztok Kosem, University of Ljubljana, Ljubljana(SI), iztok.kosem@ff.uni-lj.si

**Report**

During her 50-day STSM at the Faculty of Arts, University of Ljubljana, Tanara Zingano Kuhn has learned and tested the method of automatic extraction of lexical data from corpora, which has been developed during the compilation of Slovene Lexical Database (Gantar et al. 2016) and has been recently made part of the compilation of Collocations Dictionary of Slovene, as well as the Dictionary of Contemporary Slovene. The goal was to adapt the method for the Portuguese language and conduct its evaluation for the purposes of compilation of dictionary of academic Portuguese.

The first third of the STSM was spent on reading the relevant literature, and discussing various aspects of automatic extraction procedure such as lemma selection, defining grammatical relations, and setting criteria for automatic extraction. The other activities were related to her corpus of academic texts - although it was planned that Tanara would already have the corpus by the start of her STSM, she decided to work on it during the STSM as our discussions have pointed to the importance of having a quality corpus for not only lexicographic analysis but also the development and evaluation of constituent parts of automatic analysis. The end result was very impressive as she managed, with the help of her colleague with IT expertise, to build a 50-million-word corpus of academic Portuguese which was balanced in terms of varieties of Portuguese and contained academic texts from six different areas. Having such a corpus is undoubtedly important for ensuring a good evaluation of automatic procedures, development of automatic tools, and of course for the quality and relevance of her PhD thesis.

The most valuable outcome of Tanara’s STSM, however, is the sketch grammar for Portuguese. The initial plan was to use one of the existing sketch grammars and focus on a sample selection of grammatical relations, but after conducting a thorough analysis of available sketch grammars, Tanara’s decided to learn Corpus Query Language (CQL) and write the sketch grammar on her own. This was far from an easy task as several theoretical and practical questions arose and had to be addressed; several of those questions were linked to the poor performance of Freeling POS-tagger for Portuguese, and some were related to the actual problems of devising grammatical relation definition based on part-of-speech tags rather than parsed information. Tanara has demonstrated a great knowledge of various characteristics of the Portuguese language and an ability to quickly learn (and implement) the advantages of CQL, as well as how to address its shortcomings. Although Tanara has already said that she intends to improve the sketch grammar, the existing version is already far better than other available sketch grammars for Portuguese and the fact that Tanara intends to make her sketch grammar available to the Sketch Engine users means that the results of this STSM will have a wide impact, not only for the lexicographic community but also for the research into the Portuguese language in general.

An important part of our discussions with Tanara, and of the procedure of automatic extraction, was related to varieties of Portuguese (Brazilian Portuguese and European Portuguese) and word variants, as well as the recent orthographic reform, highly relevant in planning any language description of contemporary Portuguese. It became clear that this will be one of the key areas that Tanara’s proposed dictionary will need to address, so much time was spent on discussing how the dictionary should represent this information. It also meant that the automatic extraction procedure needed to be adapted – for Slovene, for example, all the data is extracted from one large reference corpus. For Portuguese, Tanara and I devised a procedure that includes automatic extraction of data (grammatical relations, collocates, examples, as well as all the relevant metadata) from each of the variety subcorpora; from the entire corpus, only grammatical relations and collocates, and their statistical information, are extracted. The information is then merged and variety alerts are automatically generated (e.g. for words found only in Brazilian Portuguese are assigned label BR, words far more common in European Portuguese than in Brazilian Portuguese are assigned label “typical PT”, alerting the lexicographers to such use and potential need for labels.

Other new additions to the automatic extraction procedure were clustering of collocates and the longest commonest match information. This required evaluation of various settings for clustering and the selection of the best setting for automatic extraction, and analysing longest commonest match output in Word sketches. The script was then developed by our programmer that extracts all this information from the corpus and adds it to the automatically extracted data before the export into the dictionary-writing system.

Due to all of the above mentioned activities, less time could be devoted to preparing GDEX configurations and testing different parameter values for automatic extraction. As far as GDEX configurations are concerned, a few were developed that gave fairly good results, although Tanara and I agreed that more development is needed; this will be part of the follow-up activities. As far as testing parameter values, we decided to set only a minimum frequency limit for collocates, mainly because parameter values for Slovene had been devised only after we had already compiled several hundred entries and had data to base our values on (Tanara did not have such dataset at her disposal).

Tanara and I have also discussed the ways in which we would like to disseminate the results of our work during the STSM. They include presenting the results of the STSM at one of the future meetings of WG3, writing a joint paper and giving the presentation at the next eLex conference (which will be co-located with the final ENeL meeting), and writing a paper for the special issue of the International Journal of Lexicography which will be dedicated to the results of WG3. Furthermore, we intend to continue our collaboration, as well as to promote the collaboration of our institutions, on developing the automatic extraction of data for Portuguese as well as improving existing tools for Portuguese (for example, Tanara has continuously shown interest in helping improve existing taggers for Portuguese).

In conclusion, I consider Tanara’s STSM a great success in terms of both results and future collaboration possibilities. During her visit, Tanara has demonstrated great knowledge in her area, willingness to learn, ability to ask (and answer) difficult theoretical and practical questions, and great awareness of various aspects of dictionary-making. I have no doubt that she will successfully complete her PhD and continue to build her academic career, and I look forward to our future collaboration.

Dr Iztok Kosem

Department of Translation

University of Ljubljana