**Host Report of Short Term Scientific Mission**

**COST STSM Reference Number:** COST-STSM-IS1305-29989

**Period:** 12/06/2016 to 26/06/2016

**Duration**: 14 days.

**COST Action:** IS1305

**STSM type:** Regular (from Czech Republic to Slovenia)

**STSM Title:** Graphical user interface for GDEX (Good Dictionary Examples) **Guest/STSM applicant:** Jan Michelfeit; Lexical Computing CZ;
**Host:** Iztok Kosem; Trojina, Institute for Applied Slovene Studies

**Report**

During his STSM at Trojina Institute, Jan Michelfeit analysed the feedback of Slovenian (Iztok Kosem) and Estonian (Kristina Koppel) GDEX developers on various aspects of developing GDEX configuration, and then implemented the main findings into the development of a new GDEX editor that would facilitate GDEX configuration design and offer a better overview over the results (of individual classifiers). Given the short duration of the STSM, it was envisaged that a more thorough testing and evaluation of the new editor will be conducted after the STSM.

On the first day, Jan presented his proposed solutions on certain aspects on the received feedback, and their implementation in the new editor. The original goal was to enable users to create a GDEX configuration file with a set of predetermined classifiers in a point-and-click interface, adjust the classifiers' parameters, group them into tiers by significance and set their weight in the total score. Our discussion, however, has shown that a point-and-click feature was not a priority, as normally the GDEX developers know the code enough to be able to edit it directly. Thus, Jan focused on presentation of GDEX classifier scores and improved user-friendliness of the comparison of two different GDEX configurations.

In the first week, Jan developed the first version of the tool so that certain features could be tested or at least commented on. Also, during that time, we were able to discuss the role of different classifiers in GDEX, and, based on certain findings by Kristina, introduce new types of classifiers into GDEX configurations. Initial testing of the editor was setup to work with the Estonian National Corpus, which is available on the Sketch Engine server. Working daily in the same office proved invaluable as continuous feedback and discussions on the editor and GDEX in general meant that certain improvements could often be implemented (and tested) immediately.

In the first few days of the second week, Jan developed the GDEX editor to the fully working version which included the latest additions agreed on during the first week of his STSM. Among the most important features are the options of limiting the search (and configuration evaluation) to a subcorpus, using shortcodes to get a breakdown of GDEX score by classifiers, setting sample size, and setting the level of deduplication. Each feature contributes to user-friendliness of the editor, but a breakdown of the GDEX score is perhaps the one that stands out the most, for several reasons. Among the most important reasons is the developer’s ability to not only see how a certain classifier scores sample sentences, but also to test whether that classifier even works, or works the way it is envisaged.

One possible shortcoming of the new editor is that it no longer offers two sets of examples offered by two different configurations, but rather provides a single set of examples (and scores), and then offers the option of reordering them by one or the other configuration score (or even classifier score). Still, the initial evaluation has shown that the approach used in the new editor is in fact in many ways better and more efficient.

Jan spent the final days of his STSM deploying the GDEX editor on the Sketch Engine website and on the Trojina server, testing it thoroughly. In addition, he prepared a brief documentation of the editor.

Jan and I have also discussed the ways in which we would like to disseminate the results of our work during the STSM. One of the main aims is to co-author a paper on good example extraction (with several other authors) 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, which will include further evaluation of the editor and, if needed, more upgrades.

In conclusion, I consider that Jan's STSM was a success, as all the aims have been achieved. I think the lexicographic community will benefit from having this editor at its disposal, hopefully at some point also independently of the Sketch Engine tool. I look forward to future collaborations with Jan, both in terms of development of lexicographic tools and writing academic papers.

Dr Iztok Kosem

Trojina, Institute for Applied Slovene Studies