**Host Report of Short Term Scientific Mission**

**COST STSM Reference Number:** COST-STSM-IS1305-29666

**Period:** 12/06/2016 to 12/07/2016

**Duration**: 30 days.

**COST Action:** IS1305

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

**STSM Title**: Automatic extraction of good dictionary examples for Estonian learner's dictionaries

**Guest/STSM applicant**: Kristina Koppel, The Institute of the Estonian Language, Tallinn, Estonia

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

**Report**

During her STSM at the Faculty of Arts, University of Ljubljana, Kristina Koppel has studied the characteristics of good dictionary examples and improved the configurations for good dictionary example extraction from the Estonian corpora, using the GDEX tool. She has also evaluated and provided feedback on the new version of the GDEX Editor, which was developed within another EneL STSM. The main goal was to utilise some of the techniques used in the Slovenian lexicography (e.g. Kosem et al. 2011; Kosem et al. 2013) and apply them to the Estonian language, with special focus on learner's dictionaries, including a dictionary of collocations currently compiled at the applicants institution.

The first half of the STSM was spent by discussing the characteristics of good examples and analysing the existing examples in the database of the dictionary of collocations. Namely, as was agreed before the STSM, Kristina prepared two training datasets: one containing the selected examples in the database, and the other one comprised of rejected or non-selected examples. In the first week in Ljubljana, instructions on which features of the two datasets we want to look into, and then scripts for extracting those features were prepared by our programmer. Kristina's analysis, combined with her experience of selecting the examples for the dictionaries, then focussed on identifying the features of the analysed examples that could inform the improved GDEX configurations. Several features were identified, especially those related to sentence length, sentence initial word (determined by word class), headword position etc.

Kristina used the existing GDEX configuration for Estonian as a point of departure for devising a new configuration. Drawing on the observations of their team when selecting examples for the dictionary of collocations, she wanted to both improve the parameters of existing classifiers and include completely new classifiers into the configurations, as well as possibly remove some of the non-relevant classifiers. She dedicated a great deal of time to testing the settings of each individual classifier and understanding the calculations behind it. She also compared the results of the new configuration(s) with the results of the old one, comparing not only the ranks of the same candidate sentence, but also GDEX scores, overall and of individual classifiers. During this process, she has demostrated great awareness and knowledge of different languages technologies (GDEX tool, but also tokenizers, taggers, etc.) as well as the knowledge of the characteristics of the Estonian language and of the needs of the users, in this case learners of Estonian as L2.

After deciding on the classifiers that would be used in the new GDEX configuration, Kristina focussed on deteriming the best weights for them. As it turned out, this was also one of the main changes that were implemented to the configuration, if compared to the previous one. Namely, the previous configuration used for extracting candidate examples from the corpus divided classifiers into hard ones representing 50% of the score, and soft ones, also representing 50% of the score – the difference between both types of classifiers was in the scoring; hard classifiers were mutually dependent (ALL conditions had to be met for a candidate sentence to receive 50% of the score) where soft classifiers were not (they were scored independently, each of them contributing the same share to 50% of the score). In the new configuration, the hard classifiers were still scored the same, but the soft classifiers were divided further into those with their own weight, and those that were grouped together and shared the same weight due to shared characteristics (e.g. number of elements in the sentence); in fact, this was very similar to the approach used in the GDEX for Slovene. The optimal weight settings were obtained through extensive testing and evaluation of the results.

The final week was spent on preparing the configurations for different word classes. This mainly involved fine-tuning parameters of existing classifiers; sometimes a new classifier was added due to the findings of the analysis on the two training datasets.

Overall, Kristina really impressed me with her hard work and skills, demonstrated during her STSM. In addition to utilising existing GDEX configurations at her disposal, she went one step further and came up with new classifiers that can further assist in identyfing good examples. In this way, she also contributed new knowledge to this area that I intend to implement and test for the purposes of identyfing good examples for Slovene.

Kristina 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, as well as to promote the collaboration of our institutions, on developing the techniques for automatic extraction of data from corpora.

In conclusion, I consider Kristina’s STSM to be a great success as the aims have been achieved. Not only that, working with Kristina, discussing various aspects of good example extraction and of the GDEX tool made me think of further improvements for the Slovenian GDEX configurations, which means that the usefulness of this STSM was very two-directional. 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