EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

SHORT TERM SCIENTIFIC MISSION (STSM) SCIENTIFIC REPORT

This report is submitted for approval by the STSM applicant to the STSM coordinator

Action number: CA16105
STSM title: A case study on learning material creation using a combination of automatic data extraction and crowdsourcing
STSM start and end date: 28/05/2017 to 02/06/2017
Grantee name: Iztok Kosem

PURPOSE OF THE STSM:

The proposed aim had two aims: firstly, to learn about development of L2 learning materials at University of Gothenburg, also with conducting an exercise in learning material creation, and secondly, to devise a case study for an overview of crowdsourcing resources and tools.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

On the first day, I attended a mini-workshop on the creation of language learning materials and automatic sentence selection for learner reference resources. Ildiko Pilan from Språkbanken presented her work on using machine learning to extract sentence candidates for language learning exercises, with an aim to develop a tool for teachers which would be used to select suitable examples for different CEFR levels. A demonstration of the tool was also given.
The second presentation was given by myself, and was focused on the automatic data extraction methods used in Slovene lexicography and language learning. I also presented our plans on how we intend to implement crowdsourcing techniques into our workflow.
The third presentation was given by Kristina Koppel from the Institute for the Estonian Language. She presented their lexicographic projects, and focussed on the automatic example extraction for the purposes of the collocations dictionary for L2 learners of Estonian.

The presentations were followed by a discussion between workshop participants on various subjects, including using machine learning for language learning applications, and using the potential of crowdsourcing in language learning.

The second day and the first half of the third day were dedicated to hands-on tutorial on machine learning, led by Ildiko Pilan, and began with a short presentation of the basics of machine learning, and an overview of existing studies and projects in the field of language learning. We have looked at some data used by Språkbanken in the development of

Funded by the Horizon 2020 Framework Programme
of the European Union

language learning materials. We re-did the analysis on the data, using the WEKA tool, to see machine learning in action. Each activity was followed by a group discussion in which we discussed in detail the ways in which this could be incorporated into the projects in Slovenia, not only in language learning settings but also in lexicography.

The rest of the STSM was dedicated to meetings with Elena Volodina and discussions about the crowdsourcing projects in language learning settings. I have presented our work with L1 learners, such as the Šolar corpus, and demonstrated the full cycle from error annotation to the creation of exercises based on the detected errors. Elena Volodina presented the SWELL project (electronic research infrastructure on Swedish learner language) which recently started at Språkbanken and is similar in nature to Šolar, but targets the analysis of L2 learner writing. We have identified several common aspects and problems of both projects, e.g. approach to digitization of written material, annotation of errors, and structure of the corpus. We have also dedicated significant amount of time to the discussion of potential issues in the use of crowdsourcing in language learning, e.g. copyright, teacher and learning motivation, perception of crowdsourcing in language learning community, what terms like explicit and implicit crowdsourcing actually mean (to us and to others) and so forth. We see all these topics fitting closely with the aims of WG1 in the Action, so detailed notes and suggestions for their inclusion into the WG1 Work plan were made. Finally, we also outlined a plan for a study that would provide a detailed review of existing literature on projects relevant for (explicit) crowdsourcing.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

The results of the STSM can be summarized as following:

- Attending and actively participating in tutorial on automatic example selection.
- Learning about Språkbanken infrastructure and existing projects
- Obtaining knowledge on machine learning
- Discussing crowdsourcing, its potential for the development of language materials, and various potential issues related to its use
- Preparing an outline of a plan for a study that will prepare a detailed review of existing literature on crowdsourcing
- Meetings with several members of Språkbanken staff, and discussions about potential collaboration

FUTURE COLLABORATIONS (if applicable)

I had a meeting with Lars Borin, the Director of Språkbanken. I presented my work in Slovenia, and he presented the SALDO project which aims to compile a semantic lexicon for Swedish. We have identified several parallel with our work on Collocations Dictionary for Slovene, and the SALDO project.

Overall, the STSM was a complete success, and beyond, as all the activities envisaged were completed, and additional knowledge and experiences were shared between myself and my host. We have identified several common interests and related projects, and intend to establish long-term collaboration in language learning field, both by exchanging existing knowledge and by jointly exploring the potential of crowdsourcing in the development of language learning materials.