SHORT TERM SCIENTIFIC MISSION (STSM) – SCIENTIFIC REPORT

The STSM applicant submits this report for approval to the STSM coordinator

Action number: CA16105
STSM title: Work on the learning and reading assistant tool LARA – Icelandic text
STSM start and end date: 25/12/2018 to 31/12/2018
Grantee name: Branislav Bedi

PURPOSE OF THE STSM

The intention of this STSM was to expand work on the Learning and Reading Assistant LARA for Icelandic. The general idea of this tool is to support learning a foreign or second language (L2) by reading a text and listening to a voice recording of the same text. Learners are expected to be false beginners, or at least have some knowledge of the target language to be able to rad the letters and understand basic grammar. During the process of reading, the meanings of words are heuristically acquired based on their context of use, and grammar is learned in chunks based on the occurrence grammatical constructions in the text. The process is normally driven by the learner's previous knowledge of a related language. The aim of the proposed STSM was therefore work on the LARA project and sketch ideas for further work regarding learning (not only) Icelandic language, as well as to enlarge the network of enetCollect and support work between different working groups, which, as will be described later, has been achieved.

Together with the team at the University of Geneva, the purpose was to create a marked-up content including audio into LARA. The capability was added during the STSM. The idea of using Icelandic as a case study has expanded for another languages, i.e. English, Farsi, and currently planning to add French. Since the focus during my STSM was on Icelandic, this report will generally describe work connected to that. Icelandic is a language, which poses nontrivial challenges at the levels of vocabulary, grammar and pronunciation; for a European, it is significantly more difficult to learn it than, e.g., French, German or Spanish, but it is far easier than e.g. Farsi or other middle eastern or Asian languages. The purpose was to choose a suitable text in Icelandic, including simple vocabulary and easy to read, that would be also appropriate for testing on beginner learners of Icelandic. The work primarily focused on the development of content and is detail described in the next section.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSM

During the STSM 25 -31 December 2018 (7 days), the following work was carried out:

- An appropriate text in Icelandic was selected: Tína fer í frí, which is a children´s story written by a Danish author Esther Skriver, which specialised herself in this kind of literature. This book belongs to the compulsory reading material for pupils (7 years old) at primary schools in Iceland. It

includes simple vocabulary, simple sentences, very few compound words that are clearly marked, i.e. orthographically written with a hyphen to enable an easier way of reading. The text is graphically designed as if in a verse-style, but it doesn´t rhyme. It only enables the reader to focus on 4-5 words in a line, which makes the general reading experience easier;
- A short literature review about second language learning (L2) pedagogy for reading was gathered;
- The text was transcribed in a text. format, segmented and tagged, and later on transferred to an HTML format;
- The segmented parts, or units, were voice recorded. The recording of the whole text took about 6 hours because some passages needed to be re-recorded;
- The tagging part of POS in the text has been an on-going process due to inconsistency, spelling mistakes, typos and accidentally left out words. The text had to be manually tagged because currently we do not posses any compatible automatic tagging system for Icelandic;
- Since Icelandic is a very inflectional language, a grammatical reference freely available online was needed to connect to LARA. One such tool BÍN with open access was found. This tool displays inflectional forms of nouns, adjectives, numbers and verbs. It was incorporated to the Icelandic text as grammar information. This functions as a hyperlink to the original website, and in particular to the specific word;
- Tasks for recording of individual words have been created. This will enable the learner to click on a loudspeaker icon of each head word in order to hear its pronunciation;
- A draft of an article was discussed and also written together with the team from the University of Geneva.
- A preliminary testing of the first LARA for Icelandic text was planned for the post STSM time and has already been carried out with on two learners of Icelandic in order to receive some immediate feedback before doing the pilot testing with real learners of Icelandic in the language course at the University of Iceland.

During the stay, the work was very intensive every day. Regular meetings and discussions took place in order to improve the work procedure and features of LARA. This STSM created the basis for my applying for a post-doc grant at the University of Iceland. I am currently working on the application.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

The main result obtained during the STSM was a completely marked-up text with voice-recorded segments that all together represented the whole Icelandic text. This has been put online; it is working and accessible for the purposes of individual case studies. Consequently, the results from case studies are being used to further improve LARA for Icelandic. Some features may also be adopted to other texts in other languages.

Here below the main results are discussed in a more detail:

- The selection of texts for LARA needs to be in concordance with the level of learners and the purpose of learning. The initial text for Icelandic has been selected because of its interesting story, simplicity, basic vocabulary and simple sentences, which is very suitable for beginning learners. This is pedagogically supported throughout the L2 literature.
- Rules for tagging a text in LARA have been developed and are constantly being improved based on examples from various text. The idea behind the tagging is to have a unified system of rules that would help serve the learning purpose when the texts are translated, i.e. not only to serve the purpose of a collection of words as in a dictionary but create translations that teach the learner about the use of phrasal verbs, idiomatic expressions and individual words within the context of a particular text.
- Segmentation of texts appears to differ from language to language and from text to text. Large passages of texts are not suitable to become large segments. When these are recorded the learner can easily loose the track. For this reason, the texts are recommended to be segmented into smaller segments that “make sense”, i.e. they include direct speech of one voice, which may

already include several shorter or longer sentences. After such instance it is then advisable to create another segment. The feeling of the content provider, who does the segmenting, is often and rightly listened to.

- Many features for LARA for Icelandic were added during the STSM. One of them was the inclusion of grammar information on forms of inflections.
- Various issues regarding programming have been resolve, but many of them remain and are a task for future work.

FUTURE COLLABORATIONS (if applicable)

A) Developing or including an automatic annotating tool for (not only) Icelandic would be very beneficial, especially if texts are going to be crowdsourced and the content provides will not be able to annotate the texts properly, or sufficiently enough in large texts.
B) The idea is to continue working with the team at the University of Geneva by applying for a post-doc grant at the University of Iceland. In case of a positive outcome, this will be very favourable for both parties and will enable to include other collaborators for, e.g., Old Norse texts and Icelandic translations.
C) We are looking for other collaborators that will help include the same for other languages.
D) We would like to present the current work on LARA during the upcoming enetCollect meeting in Lisbon (13-14 March 2019) and ask for feedback as well as to address the audience with a question to collaborate.

I received the above report and I approve it.

Best regards,

Dr. Emmanuel Rayner

University of Geneva