Report on the outcomes of a Short-Term Scientific Mission[1]

Action number: 18209

Grantee name: Ilan Kernerman

Details of the STSM

Title: **Linking lexicographic resources and CEFR-graded vocabulary lists**

Start and end date: 03/04/2022 to 12/04/2022

Description of the work carried out during the STSM

Description of the activities carried out during the STSM. Any deviations from the initial working plan shall also be described in this section.

The objective of the STSM was to examine and plan the linking of lexicographic data to difficulty graded vocabularies, with the goal of enhancing the development of language teaching and training materials.

From April 4 to 12, I had daily meetings with my Dutch Language Institute (INT) hosts, Kris Heylen and Carole Tiberius, occasionally joined by their colleagues and by K Dictionaries members (via zoom) in some presentations and discussions.

We began by studying language learning vocabulary lists, proficiency levels and needs, and noted in particular the Common European Framework of Reference for language learning, teaching and assessment (CEFR) with its six-level grading system (A1-A2-B1-B2-C1-C2, [1]). We found CEFR-corresponding vocabulary lists for 14 languages [2], developed during the Kelly [3, 4] and CEFRLex [5]) projects as well as with further support of the Council of Europe [5]. Of these language lists, only Dutch (NT2Lex, [6, 7, 8]) turned out to have sense disambiguation, linked to Open Dutch Wordnet [9, 10]. It should be noted that sense linking is available in CEFR related English resources, in Cambridge University Press's English Profile and *Oxford Advanced Learner's Dictionary* 10/e, as well as by Pons for German, but these are not open. We have thus decided to focus on the Dutch list, and in fact used Cornetto, which is a non-open version of Open Dutch Wordnet, and which overlaps with NT2Lex to some extent but not entirely.

Our main research questions concerned how to link a polysemous CEFR-graded word to the appropriate sense in a lexicographic entry (and deal with potential occurrence of other senses of that

[1] This report is submitted by the grantee to the Action MC for approval and for claiming payment of the awarded grant. The Grant Awarding Coordinator coordinates the evaluation of this report on behalf of the Action MC and instructs the GH for payment of the Grant.

word in higher CEFR levels), detect higher-graded CEFR words in a lower level's word definition or example (and likewise in translations), and distinguish production and reception issues. There were other questions related to ownership and license issues regarding the use of resources in the LLOD cloud, and we also started planning a new pan-European project with multidisciplinary partners for the convergence of lexicographic data with graded proficiency lists.

Once the main issues were laid out, we prepared a preliminary test to match a small list of A1-level Dutch words from NT2Lex/Cornetto to our existing resources. We asked the KD office for the corresponding data of Global entries, which are part of a cross-lingual series that includes learner's definitions and usage examples (as well as multilingual translations) and thus seemed better suited to offer CEFR matching lexicographic descriptions than those available from Open Dutch WordNet or INT's own ANW (*Algemeen Nederlands Woordenboe*k [Dictionary of Contemporary Dutch]). Then we proceeded to tag each word/sense in the Global entry to its CEFR counterpart, compared the senses with Open Dutch WordNet and ANW, reviewed the fields and topics presented in the original CEFR reference level descriptions [6] that served to develop the current lists, and checked the equivalence of the translations. More thought will be needed to deal with higher grading levels and other languages, and to resolve sense division of CEFR vocabulary.

Our final task was to match the CEFR-NL word list with the Global, NT2Lex/Cornetto and ANW ones, explore the differences and evaluate their numbers and proportions, and get a glimpse of the challenges facing us with future adjustment of the CEFR resource.

Description of the STSM main achievements and planned follow-up activities

Description and assessment of whether the STSM achieved its planned goals and expected outcomes, including specific contribution to Action objective and deliverables, or publications resulting from the STSM. Agreed plans for future follow-up collaborations shall also be described in this section.

The STSM has largely achieved its goals and outcomes, including contribution to the Action and planning a publication and follow-up collaboration, mainly:

- To study best methods to enhance lexicographic information on the difficulty of words and phrases for different learner levels – we decided to focus on the CEFR grading system and identified major challenges.
- To explore how to integrate such information with language learning applications – we made first tests on linking lexicographic resources with a CEFR list in preparation of incorporating such merged information in language learning applications.
- To plan the upload of by-products to the Linguistic Linked Open Data cloud – we began studying the legal implications of offering privately-owned data on the LLOD cloud to find the most appropriate form of license.
- To contribute to NexusLinguarum CA – by the means of a comprehensive assessment of the feasibility and usefulness of the above-described linking, including multilingual applications and eventual upload to the LLOD cloud.
- To plan a joint paper for presentation – at the Nexus-supported conference 'LLOD approaches for language data research and management', due in Vilnius on September 21-22, and for subsequent publication in the journal *Rasprave*.
- To prepare the ground for a wide-range project on this topic – bringing together academia-industry lexicographic, language teaching, linked data and knowledge communities, in the framework of Horizon Europe or Digital Europe programs starting in 2023.

The annexes present some of the basic initial results of the work carried out in the STSM.

References

[1] CEFR. Council of Europe. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. 2001. https://www.coe.int/en/web/portfolio/the-common-european-framework-of-reference-for-languages-learning-teaching-assessment-cefr-?_82_struts_action=%2Flanguage%2Fview&_82_languageId=fr_FR

[2] Arabic, Chinese, Dutch, English, Estonian, French, German, Greek, Norwegian, Polish, Russian, Slovenian, Spanish, Swedish.

[3] Kelly. 2009-2011. https://spraakbanken.gu.se/en/projects/kelly

[4] Kilgarriff, A. et al. 2014. 'Corpus-Based Vocabulary Lists for Language Learners for Nine Languages'. *Language Resources and Evaluation*, 48.1, 121–63. https://link.springer.com/content/pdf/10.1007%2Fs10579-013-9251-2.pdf

[5] CEFRLex. 2014-2021. https://cental.uclouvain.be/cefrlex/.

[6] RLDs. (CEFR) Reference level descriptions developed so far, 2022. https://www.coe.int/en/web/common-european-framework-reference-languages/reference-level-descriptions-rlds-developed-so-far.

[7] GR4L2. Building CEFR-graded resources for foreign and second language learning. https://uclouvain.be/fr/instituts-recherche/ilc/plin/gr4l2.html. 7 Dec 2021

[8] NT2Lex. https://cental.uclouvain.be/nt2lex/

[9] Tack, Anaïs, et al. 2018. 'NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet'. T*he 13th Workshop on Innovative Use of NLP for Building Educational Applications* (NAACL 2018 Workshops), https://www.aclweb.org/anthology/W18-0514/

[10] Open Dutch WordNet. http://dutchframenet.nl/open-dutch-wordnet/