Japan Association for English Corpus Studies | SIG Corpora and the CEFR

Spring Forum | 13 May 2023

Linking lexicographic and CEFR resources

Kris Heylen[1], Ilan Kernerman[2], Carole Tiberius[1]
[1]Dutch Language Institute, [2]Lexicala by K Dictionaries

Abstract

The Common European Framework of Reference for Languages (CEFR) promotes the development of empirically based datasets for 30 languages of Europe according to graded proficiency levels, by grading word difficulty prevalent in native and additional language learning, in production and reception tasks, for text readability analysis and vocabulary testing. We will present ongoing work to better support the creation of relevant teaching and training materials by linking CEFR difficulty-graded word lists with lexicographic data, cross-lingualizing different language sets, and uploading the by-products to the Linguistic Linked Open Data cloud. The case studies involve, on the one hand, the corpus-derived CEFR-graded resources compiled for a number of languages in the CEFRLex project and, on the other hand, the lexicographic datasets for Dutch of K Dictionaries and the Dutch Language Institute. The challenges we discuss include (a) evaluating the words in CEFR lists and linking their appropriate senses in dictionaries, (b) ensuring the additional lexicographic components exclude words from higher CEFR levels, and (c) creating CEFR lists for languages that don't have them yet and linking them to lexicographic data. Additionally, we seek diverse partners with language learning expertise, lexicographic resources, and linked data know-how, to join in a future project.

Keywords
language learning; proficiency levels; CEFR lists; lexicographic resources, LLOD cloud