▶ Linking Lexicographic Resources to Language Proficiency-Level Applications

Kris Heylen,
Dutch Language Institute, Netherlands, kris.heylen@ivdnt.nl
Ilan Kernerman,
K Dictionaries – Lexicala, Israel, ilan@kdictionaries.com
Carole Tiberius,
Dutch Language Institute, Netherlands, carole.tiberius@ivdnt.nl

Purpose: We aim to enhance the development of vocabulary teaching and training materials by converging difficulty-graded word lists with lexicographic data. Grading word difficulty is prevalent in both native and additional language learning, in production and reception tasks, and for text readability analysis and vocabulary testing. Our objectives are to upgrade the usability of such resources for creators of vocabulary learning materials – by enriching them with semantic information such as definitions, examples of usage, and multiword expressions (and possibly more) from dictionaries – cross-lingualize the different language sets, and upload the by-products to the Linguistic Linked Open Data cloud.

Design/methodology/approach: 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. Each of the six CEFR levels – from beginner, A1, through A2, B1, B2, C1, to advanced C2 – refers to specific situations and conditions, and includes corresponding vocabulary in every language. We will associate pedagogical and multilingual lexicographic data with the words in the CEFR lists with the aid of smart matching and linking techniques and monolingual and multilingual sense alignment methods.

Findings: Since their introduction in the early 2000s, CEFR graded lists have been developed for approximately 15 languages, with most lists having gradings only on the lemma level. Only English has comprehensive lists (with Cambridge and Oxford advanced learner's dictionaries) with proficiency gradings on the sense level, but the resulting data is not available for other open applications. The list for Dutch has been linked in a probabilistic way to Dutch WordNet synsets but, as a consequence, it also is quite noisy.

Research limitations/implications: The primary drawback of most existing CEFR lists is that they do not disambiguate polysemous words. Secondly, when linking them to their corresponding dictionary components, it is necessary to assure that the words used within definitions, examples, and expressions are not situated on higher CEFR levels, and likewise for their equivalents in the other languages.

Practical implications: The challenges are to (a) evaluate the words in existing CEFR lists and link their appropriate senses in dictionaries, (b) make sure the additional lexicographic components contain no words from higher levels, (c) create CEFR lists for languages that do not have them yet and link them to lexicographic data, and (d) find an appropriate project framework and a range of competent partners with language learning expertise, lexicographic resources, and link data know-how.

Originality/Value: Previous CEFR lists projects, including Kelly (2009) and CEFR-Lex (2017), as well as a few carried out individually (e.g. Estonian), have had differing results. Our project will use their relevant achievements while gradually expanding the scope to all CEFR languages and attending to all the issues described above.

Keywords: *language learning, proficiency levels, CEFR lists, lexicographic resources, LLOD cloud.*

Research type: Conceptual paper.