Review of: "A Corpus Analysis of Polysemy in CEFR-based English Textbooks"

Raphael Berthele[1]

1 University of Fribourg

Potential competing interests: No potential competing interests to declare.

Thoughts on the paper «A Corpus Analysis of Polysemy in CEFR-based English Textbooks»

The paper explores the linguistic, more specifically some lexical characteristics, of teaching materials widely used in English as a foreign language teaching. As the author rightly points out, these materials, in particular if the target language is not widely used in the learners' everyday environment, are an important source of input for the learners. On the other hand, it is clear that one should not simplistically equate the teaching materials with the learners' input, as there is obviously teacher and peer input in class and most likely also input from other, written or audiovisual sources.

Hicham Lahlou is interested in two main questions: First, to what extent the most frequent lemmas in the two textbooks overlap (they are used for different levels of English), and to what extent these high-frequency lemmas in the textbooks are used with their different senses as listed in the WordNet database.

The author doesn't explicitly spell it out, but when he says 'words', he mostly means content words. As many grammatical words (e.g. determiners, prepositions, complementizers) have very high corpus frequencies, they would be shared by the two textbooks anyway. However, there are also verbs that are grammatical words, as for example 'be', which is in Hicham Lahlou's list. It was not clear to me whether the selection of target words mechanically based on the part of speech is really the best option for what the author wants to find out. The polysemy of 'be' is taken into account here, but how about the notoriously polysemous preposition 'over'? I don't have a simple answer as to how this selection problem could be solved, as from a cognitive linguistic point of view (which is the one adopted by the author and shared by myself), there is no strict separation between grammatical words and content words anyway.

I have the feeling that the author was surprised to discover a large extent of overlap between the 100 most frequent content lemmas in the two textbooks respectively. If we think of the Zipfian distribution that words in natural language exhibit, I don't think this is so surprising: The 100 most frequent content lemmas are all extremely high frequency and are therefore rather likely to occur in any English text. As the author rightfully points out, it makes sense to start teaching the most frequent words first, in order to rapidly increase text coverage. The 100 most frequent content words are just the beginning of this, but obviously a valuable one.

I liked the method to look up the senses of these lemmas in WordNet. Maybe the author could provide a tutorial for teachers how to go about, with useful links to the tools he used? It seems to me that it would be interesting if teachers could do their own simple analyses of their teaching materials, along the lines proposed by the author, and explore the words and senses they want to teach.

The author considers the teaching of polysemy a good thing. I certainly agree that polysemy is an important part of vocabulary knowledge. What I was wondering was the extent to which exposure to rich polysemy from the onset is a good idea: From usage-based models of learning we know that the level of contingency of form and function/meaning is a crucial aspect of the learnability of new constructions (be they morphemes, words, or larger units). If we start by teaching richly polysemous words right from the onset, it is unlikely that learners will pick them up just by being exposed to them. Even if they get explicit instruction of such lemmas, it might be just too much information at once on a single lexical item, as there are too many associations from senses to the lexeme. As is the case in general with variability in learning, polysemy makes learning more complicated and slower. However, if it is an important feature of the target domain to be learnt, there is no way around mastering it. As the author rightly points out, polysemy is part of the target language's semantics, and ultimately a proficient learner needs to master it. Maybe by extendingthe methodology proposed in this paper we could investigate not just the frequency of words, but also the frequency of their different senses and start by teaching the most frequent senses of the most frequent words first.