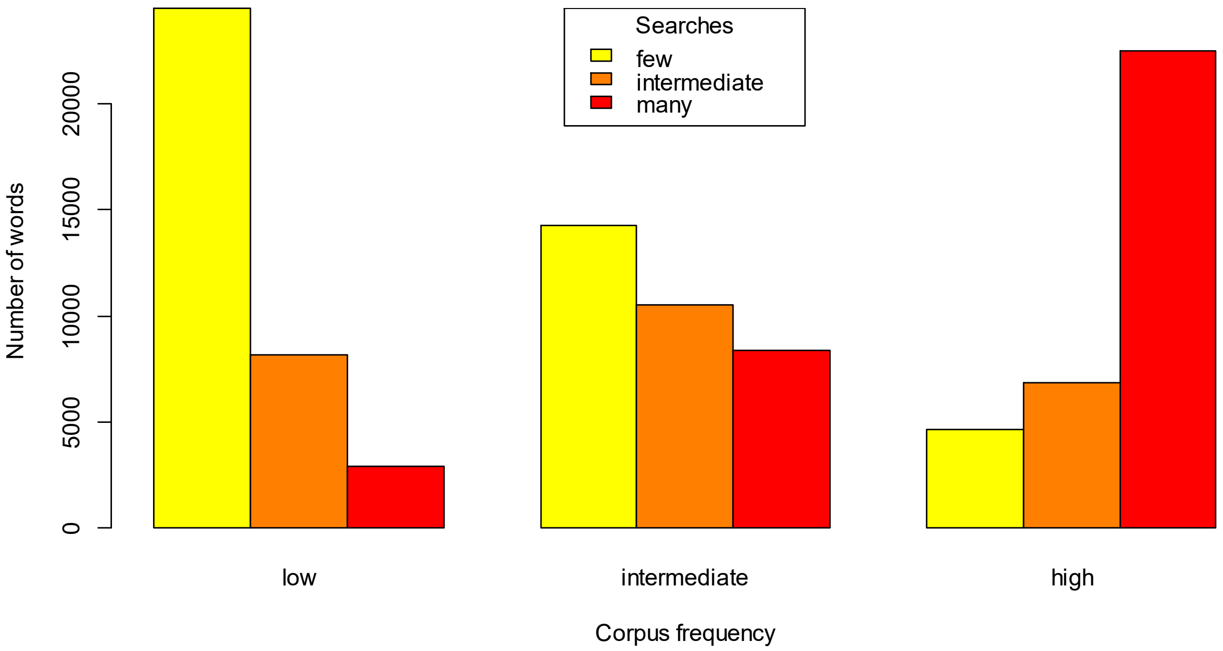
Lexical exponents predicting English Wiktionary consultations

Making a good dictionary is a huge undertaking, taking a team of experts (called lexicographers) at least a few years of intense work, even decades for more ambitious projects. Contrary to popular belief, it is far from straightforward to tell how many and which words belong to a language, so any lexicographic project faces difficult choices on which words to include, and which words to exclude.

A dictionary will be of maximum utility to its users if it covers the words that they are likely to look up. How can a lexicographer know which words will be popular? If a dictionary has a record of user activity, then we can count how many times any word has been looked up. We can then try to figure out what properties of words make them popular (or not) with dictionary users: what makes certain words particularly salient.

Recent research suggests that *lexical frequency*, that is how often a word appears in written and spoken texts, is one such factor, with the more frequent words being more likely to be consulted. This effect is evident in the following figure (De Schryver, Wolfer and Lew 2019, http://dx.doi.org/10.17576/gema-2019-1904-01):



In this project, we also plan to explore at least three further factors:

(1) the average age at which a given word is learned by native-speaking children (the technical term is *age of acquisition*);

(2) the extent to which a word is known to adult native speakers (*lexical prevalence*); and

(3) how many distinct meanings (or *senses*) a word has (technically known as *degree of polysemy*).

To be able to do that, we need to obtain and then combine a number of lexical datasets (basically, these are long word lists with specific information attached to the words). First, we plan to acquire extensive logs of user look-ups in the English Wiktionary, and process them to extract consultation frequency information for all Wiktionary headwords: this will give us information on which words users like to look up and how often. Moving on to potential predicting factors, lexical frequency will be established by building frequency lists from very large contemporary collections of English texts (called corpora), or re-using existing frequency lists. Useful data on age of acquisition and lexical prevalence have very recently been published as a result of research. Finally, for the number of meanings we can count the senses in the existing Wiktionary entries.

All these data need to be linked. Once this is done, we will build mathematical models that will try to predict, in the best possible way, the popularity of a word from its lexical characteristics. For this, we plan to use a number of advanced modelling techniques. The resulting models will then tell dictionary makers which lexical items should be prioritized in lexicographic work, so they can make more useful dictionaries more quickly. It will also be theoretically interesting to know what makes dictionary users look up words, because it tells us something about how the language in our minds works.