Sascha Wolfer/Robert Lew

PREDICTING ENGLISH WIKTIONARY CONSULTATIONS

Keywords Wiktionary; English; corpora; frequency; prevalence; age of acquisition; polysemy

Dictionaries have been part and parcel of literate societies for many centuries. They assist in communication, particularly across different languages, to aid in understanding, creating, and translating texts. Communication problems arise whenever a native speaker of one language comes into contact with a speaker of another language. At the same time, English has established itself as a *lingua franca* of international communication. This marked tendency gives lexicography of English a particular significance, as English dictionaries are used intensively and extensively by huge numbers of people worldwide.

In doing so, users make choices about which words to look up, and our aim is to identify the lexical variables that affect the likelihood of those choices by using the log files of a popular crowd-sourced dictionary: the English Wiktionary. The choice of the English Wiktionary is motivated by the availability and size of the log files. While not seeking a single lexical processing or representation model, we are interested in what drives people's decisions to look up a specific word in terms of language experience. We are contributing towards at least two research questions: 1) What makes people interested in specific words, what prompts them to seek information on these words in lexical resources? 2) Can we formulate guidelines for dictionary compilation by offering empirically-based quantifiable measures of which specific words are more likely to be sought by users, so lexicographic work can prioritize these words, and words exhibiting similar characteristics? We note that the specific look-up context as well as idiosyncratic user characteristics cannot be known at the stage of lexicographic design, and so our approach also ignores these factors.

One factor that is already known to guide people's look-up behaviour is corpus-based lexical frequency. While, quite surprisingly, the positive relationship between dictionary look-up and corpus frequency did not turn out to be apparent at all in initial studies (De Schryver/Joffe 2004; De Schryver et al. 2006; Verlinde/Binon 2010), it has since been established empirically with some confidence (Koplenig/Meyer/Müller-Spitzer 2014; Müller-Spitzer/Wolfer/Koplenig 2015; De Schryver/Wolfer/Lew 2019). However, we see a clear advantage in including further variables. Metrics reflecting other properties of words (some of them closely related – but not identical – to corpus frequency) can help us understand better the effect of corpus frequency and the relationships between different variables predicting look-up behaviour. As a first step, we will consider word prevalence, age of acquisition, and number of senses (or degree of polysemy) of the headword.

The prevalence of a word is the extent to which it is known amongst the native-speaking population. It stands to reason that words which occur with relatively higher frequency in texts and discourse should be more likely to be known by a large proportion of the speakers (Weizman/Snow 2001; Longobardi et al. 2015). However, it remains to be seen whether – with the effect of frequency controlled for – word prevalence still is a relevant predictor of look-up frequency, and if so, in which direction.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Age of acquisition is the age at which a word is, on average, acquired by native speakers in the process of (naturalistic) L1 acquisition. One might expect that this could play a role in how words acquired earlier, possibly being more deeply entrenched in the mental lexicon, get to be looked up. Age of acquisition has been found to have important and long-lasting effects on language behaviour (Ellis/Lambon Ralph 2000; Garlock/Walley/Metsala 2001; Juhasz 2005; Kuperman/Stadthagen-Gonzalez/Brysbaert 2012).

The concept of word sense is not without problems (Kilgarriff 1997; Hanks 2000) and there has been a long-drawn-out debate about the boundaries between polysemy and homonymy. To steer clear of the essentialist debate of whether words 'have' senses, we adopt a pragmatic approach of considering *lexicographic* senses: the separate blocks of meaning description as given in a dictionary. Degree of polysemy is, then, operationalized as the number of dictionary senses in the English Wiktionary itself. We have known for about 70 years (Zipf 1949) that the more frequent words tend to have more senses. However, the degree of polysemy may hold predictive potential above and beyond that of mere word frequency (Müller-Spitzer/Wolfer/Koplenig 2015).

Lexical frequency, prevalence, age of acquisition, and degree of polysemy will obviously not explain all of the variance in our model. As in any statistical model, there will inevitably remain unexplained variation represented by *residual* or *error* variance. Detailed investigation of this residual variance (complemented also with a more qualitative perspective on the observed data), additional factors might have to be brought into the picture to more fully account for look-up behaviour.

With our analyses, we hope to provide more information on the lexical variables that affect look-up behaviour – apart from mere corpus frequency. At the same time, we will try to shed some light on (groups of) headwords that might not follow the overall pattern of the data. Such outliers could point to the fact that some additional variables (or interactions between variables) have to be taken into consideration to broaden our understanding of how people use dictionaries.

References

De Schryver, G.-M. et al. (2006): Do dictionary users really look up frequent words? – On the overestimation of the value of corpus-based lexicography. In: Lexikos 16, pp. 67–83.

De Schryver, G.-M./Joffe, D. (2004): On how electronic dictionaries are really used. In: Williams, G./Vessier, S. (eds.): Proceedings of the Eleventh EURALEX International Congress (EURALEX 2004), Lorient, France, July 6–10, 2004, Vol. 1. Lorient, pp. 187–196.

De Schryver, G.-M./Wolfer, S./Lew, R. (2019): The relationship between dictionary look-up frequency and corpus frequency revisited: a log-file analysis of a decade of user interaction with a Swahili-English dictionary. In: GEMA Online Journal of Language Studies 19 (4), pp. 1–27. http://doi.org/10.17576/gema-2019-1904-01.

Ellis, A. W./Lambon Ralph, M. A. (2000): Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: insights from connectionist networks. In: Journal of Experimental Psychology: Learning Memory and Cognition 26 (5), pp. 1103–1123. http://doi.org/10.1037/0278-7393.26.5.1103.

Garlock, V. M./Walley, A. C./Metsala, J. L. (2001): Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults. In: Journal of Memory and Language 45 (3), pp. 468–492. http://doi.org/10.1006/jmla.2000.2784.

Hanks, P. (2000): Do word meanings exist? In: Computers and the Humanities 34 (1–2), pp. 205–215.

Juhasz, B. J. (2005): Age-of-acquisition effects in word and picture identification. In: Psychological Bulletin 131 (5), pp. 684–712. http://doi.org/10.1037/0033-2909.131.5.684.

Kilgarriff, A. (1997): I don't believe in word senses. In: Computers and the Humanities 31 (2), pp. 91–113.

Koplenig, A./Meyer, P./Müller-Spitzer, C. (2014): Dictionary users do look up frequent words. A log file analysis. In: Müller-Spitzer, C. (ed.): Using online dictionaries. (= Lexicographica Series Maior 145). Berlin, pp. 229–249.

Kuperman, V./Stadthagen-Gonzalez, H./Brysbaert, M. (2012): Age-of-acquisition ratings for 30,000 English words. In: Behavior Research Methods 44 (4), pp. 978–990. http://doi.org/10.3758/s13428-012-0210-4.

Longobardi, E. et al. (2015): Children's acquisition of nouns and verbs in Italian: contrasting the roles of frequency and positional salience in maternal language. In: Journal of Child Language 42 (1), pp. 95–121. http://doi.org/10.1017/S0305000913000597.

Müller-Spitzer, C./Wolfer, S./Koplenig, A. (2015): Observing online dictionary users: studies using Wiktionary log files. In: International Journal of Lexicography 28, pp. 1–26. http://doi.org/10.1093/ijl/ecu029.

Verlinde, S./Binon, J. (2010): Monitoring dictionary use in the electronic age. In: Dykstra, A./Schoonheim, T. (eds.): Proceedings of the XIV Euralex International Congress. Ljouwert, pp. 1144–1151.

Weizman, Z. O./Snow, C. E. (2001): Lexical input as related to children's vocabulary acquisition: effects of sophisticated exposure and support for meaning. In: Developmental Psychology 37 (2), pp. 265–279. http://doi.org/10.1037/0012-1649.37.2.265.

Zipf, G. K. (1949): Human behavior and the principle of least effort. Cambridge, MA.

Contact information

Sascha Wolfer
Leibniz-Institut für Deutsche Sprache
wolfer@ids-mannheim.de

Robert Lew
Adam Mickiewicz University
rlew@amu.edu.pl

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.