Word frequency calibration using external information and topic models

Rastislav Hronský (R.Hronsky@tilburguniversity.edu)[1], Emmanuel Keuleers[1]
[1]Department of Computational Cognitive Science, Tilburg University

Word frequency is one of the most powerful predictors of behavioral responses in language processing tasks, but it is neither stable nor invariant. Outside the high-frequency range, frequencies are highly corpus-dependent: a well-known example is that frequencies derived from TV subtitles explain more variance in lexical decision latencies than frequencies from much larger, carefully curated corpora [3]. Several recent methods address this context-dependence by selecting or weighting corpus contexts through fitting to reference behavioral data [5, 6]. While these methods produce improved frequency estimates, a limitation is that optimization targets the same type of data, typically lexical decision, against which the resulting frequencies are then validated, raising circularity concerns. We address this by calibrating corpus context using independent external 'seed' information. Using LDA [2] to identify topics in a corpus, we searched for topic-mixing proportions that best match two types of external seeds: the set of 10 most-visited Wikipedia pages of 2025 [1], and word association norms from the Small World of Words project [4] (SWOW). We then evaluated how the resulting topic-weighted frequencies fit lexical decision latencies. Although word associations are themselves behavioral data, they do not involve accuracy or reaction time, the two measures that frequency estimates are typically validated against.

Methodology. We used two English corpora in two parallel simulations: Wikipedia articles [10] (approx. 509MB) and movie subtitles [8] (approx. 466MB). Each corpus was decomposed with LDA using a range of settings for the number of topics and a word vocabulary adopted from the set of lexical stimuli used in the British lexicon project (BLP) [7]. Two reference values of R^2 were then calculated: baseline and skyline. The baseline reflects the fit of word frequencies, as derived from a corpus in the customary way (by counting word tokens), to the lexical decision (LD) latencies from the BLP. The skyline reflects the fit of a particular set of corpus topics weighted to maximize the fitness to LD data, such that, without changing the topics' word distributions, it should not be possible to find a better fitting mixture. We then compared the topic mixtures calibrated using the two seeds. To calibrate a topic mixture, we simply used word frequencies derived from the seed as a dependent variable in a linear model, the predictors being the candidate topics, and then taking the coefficients as the calibrated topic proportions. For the Topviews seed, we tested each of the 10 most-visited pages; for the SWOW, we tested random subsets of its full set of words of increasing sizes (32-full size) in 30 runs. The code can be found at github.com/hrasto/amlap26-word-frequency-calibration.

Results and conclusion. Our simulations show a clear difference between seed types. Wikipedia page seeds improved fits, but not as consistently as SWOW did. For reasons that are not yet clear to us, pages discussing movies provided the best results. SWOW word associations worked better: the number of word types improved fits monotonically, starting to plateau and approach the skyline R^2 with 128 to 512 word types used as a seed (the final proximity to the skyline was corpus dependent, see Figure 1). Our results suggest that corpus frequencies can be calibrated using small, easily obtainable seeds. While our current results only show that "general" calibration is possible, future research could explore whether our method allows frequency norms tailored to be calibrated to specific age ranges, occupational registers, or social contexts.

References

[1] Wikipedia topviews. https://pageviews.wmcloud.org/topviews. Accessed: 2026-05-08.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[3] Marc Brysbaert and Boris New. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990, 2009.

[4] Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, 51(3):987–1006, 2019.

[5] Rastislav Hronskỳ and Emmanuel Keuleers. Word probability re-estimation using topic modeling and lexical decision data. In *Annual Conference of the Cognitive Science Society*, pages 188–194, 2021.

[6] Brendan T Johns, Michael N Jones, and DJK Mewhort. Using experiential optimization to build lexical representations. *Psychonomic Bulletin & Review*, 26(1):103–126, 2019.

[7] Emmanuel Keuleers, Paula Lacey, Kathleen Rastle, and Marc Brysbaert. The british lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic english words. *Behavior research methods*, 44(1):287–304, 2012.

[8] Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, 2016. European Language Resources Association (ELRA).

[9] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[10] Wikimedia Foundation. Wikipedia (english), snapshot 20231101. Hugging Face Datasets (wikimedia/wikipedia, config 20231101.en), 2023. URL https://huggingface.co/datasets/wikimedia/wikipedia.

Figures and Tables

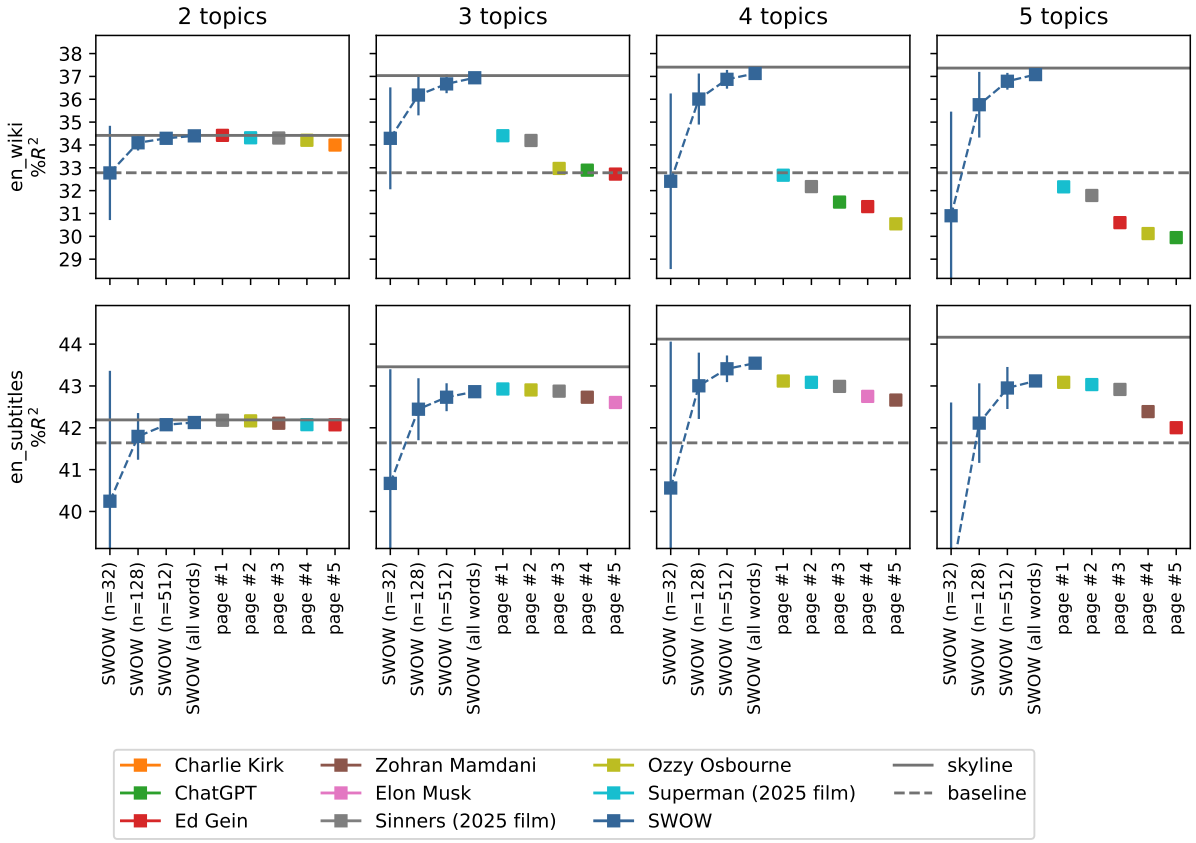

Figure 1: Overview of the results. The top panel shows results for the Wikipedia corpus and the bottom panel the subtitle-based corpus. On the left side of each plot, the connected points depict the effect of the SWOW seed, as the number of included word types in the simulation grows (the whiskers represent the standard deviation calculated over 30 random subsets, the points represent the mean). On the right side of each plot, there are five points depicting the effect of the five most potent Topviews seeds. The baseline (dashed horizontal line) shows the R^2 obtained with original corpus frequencies; the skyline (solid horizontal line) shows the R^2 obtained with the topic mixture calibrated on the LD data themselves.

Optional Supplemental Information

LDA. LDA is a generative, bag-of-words model of digital text, where every document is modeled as a mixture of topics, and every topic has a unique word-probability distribution associated with it. A component of a fitted LDA model is then a matrix of topic-word importance scores, of the shape `(number_of_topics, vocabulary_size)`. To obtain this matrix, we used the `.components_` attribute of a fitted LDA model in scikit-learn[9], and we normalized it row-wise, turning every row into a probability distribution.