Are relative frequencies really independent of corpus size?

One of the most basal corpus-linguistic measures, but one that is used far beyond corpus linguistics itself, is the relative (or normalized) frequency of lexical units. The idea is quite simple: whenever we want to compare the frequency of (for example) a word between two corpora of different size, we divide the raw frequency of the word by the corpus size and multiply the result by a normalization quantity to get a measure like *x* tokens in 1 million tokens. The goal is clear: we want to obtain a frequency measure that is independent of the size of the underlying corpus. However, it is a well-known fact in corpus linguistics that most, if not all, quantities computed based on corpora vary systematically with corpus size (Baayen 2001; Tweedie & Baayen 1998; Koplenig & Wolfer & Müller-Spitzer 2019). In our contribution, we want to test if this also true for relative frequencies and – if this turns out to be the case – how we could counteract biases and problems.

For this purpose, we use comparable (non-parallel) news corpora for, currently, seven languages (Arabic, English, Finnish, French, German, Latvian, and Vietnamese)[1] from the website of the "Wortschatz Leipzig"[2]. For a corpus with one million shuffled sentences per language, we vary the corpus size in 20 consecutive steps from 50,000 to 1 million sentences. Each step incorporates the sentences from the previous step. In each step we calculate the normalized frequency for all word types which yields 20 frequency lists per language. Now, we can compare the normed frequencies for all word types that appear in the smallest corpus to the ones in the larger corpora. If relative frequencies hold what they promise, given a pair of two different word types w_1 and w_2, the ranking of these two types should be in agreement in both corpora, i.e. the pair should be concordant: if w_1 has a higher relative frequency than w_2 in the smaller corpus, then w_1 should also have a higher relative frequency than w_2 in the larger corpus. To quantify this logic for all available pairs of word types, we use Kendall's correlation coefficient τ as a test statistic as it directly results from the number of concordant and discordant pairs. τ should be (very close to) 1 if relative frequencies indeed control for corpus size. Moreover, τ should not vary systematically with the difference in size of the two corpora being compared.

Figure 1 shows that this is clearly not the case. τ drops systematically with difference in corpus size. Also, the overall value of τ differentiates between the languages currently in our dataset. The results for Kendall's τ can be replicated with other measures, e.g. Spearman's rank correlation coefficient or the percentage of words with a higher relative frequency in the smaller corpus.[3] With Figure 2 we try to get closer to the origin of this phenomenon. Here we gradually increase the number of types from the top of the frequency list to be used in the calculation of τ. The comparison pair is always a corpus with 50,000 sentences and one with 100,000 sentences (in Figure 1 this is the comparison at $x = 2$). τ decreases as more and more low-frequency types are included. Relative frequencies thus appear to be contaminated by the underlying corpus size, especially for these low-frequency types.

If relative frequencies indeed depend systematically on corpus size, this would be a major challenge for a variety of comparative corpus-linguistic studies, including contrastive studies between corpora of different languages. In our contribution we will elaborate on our results, try to give an explanation, and evaluate a possible solution to this challenge.

[1] Where available, we used the 2010 news corpora. For Finnish and Vietnamese, we used the 2020 news corpora. For Latvian, we used the 2011 newscrawl corpora.

[2] https://wortschatz.uni-leipzig.de/de/download [last access: 2023-01-05]

[3] If corpus size would not influence relative frequency, this measure should be (very close to) 50 % which would indicate random fluctuation. We can show that the measure is considerably higher than 50 % for all comparisons.

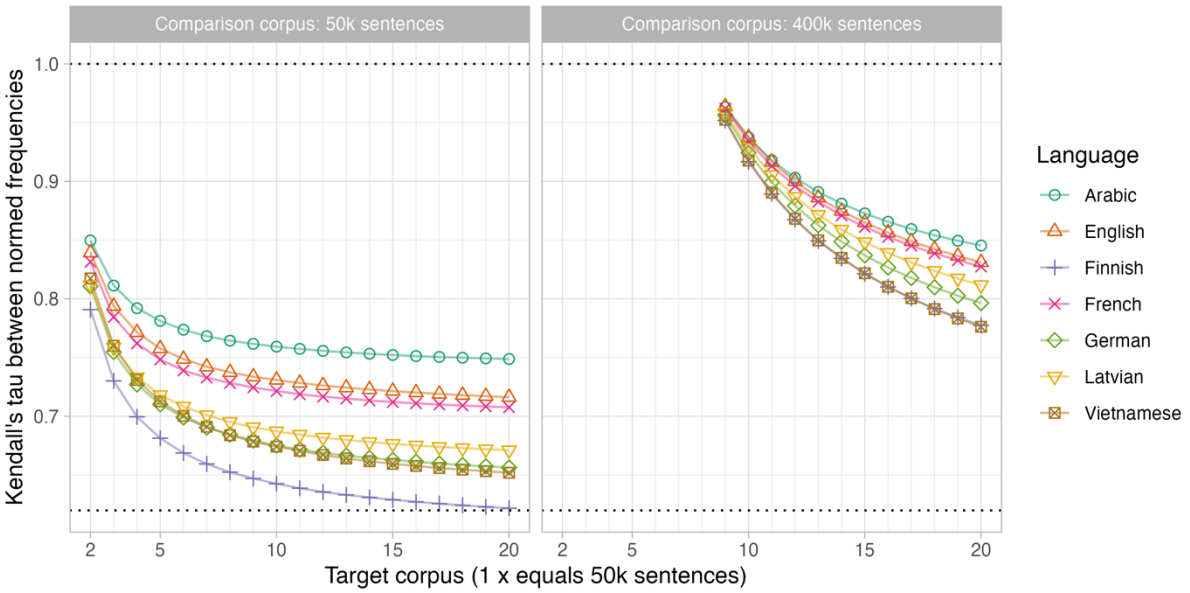

Figure 1: Kendall's τ for normed frequencies in a smaller news corpus vs. those in a larger corpus for seven languages. On the left, the smaller corpus contains 50,000 sentences, on the right 400,000 sentences. The upper dotted line represents the optimal value of $\tau = 1$, the lower dotted line indicates the minimum value in the current dataset (comparison between 50k and 1 million sentences, $x = 20$, for the Finnish corpus): $\tau = .620$.

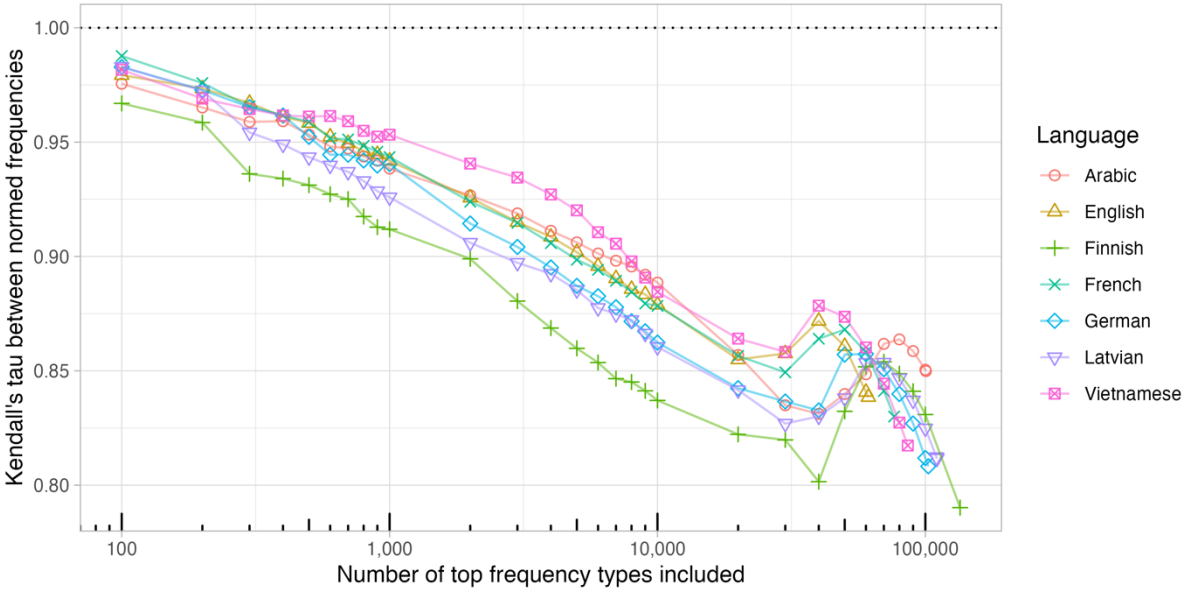

Figure 2: Kendall's τ for normed frequencies in a corpus with 50,000 sentences compared to a corpus with 100,000 sentences in seven languages and for a varying amount of top frequency types (x-axis, log-transformed). The dotted line represents the optimal value of $\tau = 1$.

References

Baayen, R. Harald. 2001. *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Koplenig, Alexander & Wolfer, Sascha & Müller-Spitzer, Carolin. 2019. Studying Lexical Dynamics and Language Change via Generalized Entropies: The Problem of Sample Size. *Entropy* 21(5). DOI: https://doi.org/10.3390/e21050464

Tweedie, Fiona J. & Baayen, R. Harald. 1998. How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities* 32(5). 323–352.