Article

Sascha Wolfer* and Alexander Koplenig

Does corpus size influence normalised frequencies?

https://doi.org/10.1515/cllt-2024-0040
Received April 11, 2024; accepted June 4, 2025; published online junio 24, 2025

Abstract: Several frequency-based measures are influenced by corpus size (e.g. lexical diversity or text similarity measures). It is largely unquestioned, however, that normalised frequencies correct for the influence of corpus size – but it has not yet been systematically tested whether and how they might be influenced by corpus size themselves. The central question is whether the normalised frequency of an element in a smaller corpus can be meaningfully compared to the normalised frequency of the same element in a larger corpus. We are testing the association between lists of normalised frequencies derived from corpus samples of different sizes from six languages. Our results suggest that the size of the underlying corpora does not negatively influence comparisons of normalised frequency lists, i.e. different corpus sizes do not lead to normalised frequencies no longer being comparable. For lower-frequency types, these associations decrease rather quickly. These empirical findings converge with predictions from statistical theory.

Keywords: corpus linguistics; corpus size; normalised frequency; relative frequency

1 Introduction

One of the most important measures in corpus linguistics is the frequency of occurrence of lexical elements. Sometimes, these are compiled into frequency lists, for example, to find "'important' words" (cf. Laurence 2021: 108), in whatever sense, in a corpus. Of course, this is not restricted to single words but can be expanded to parts of words or multi-word units.

***Corresponding author: Sascha Wolfer**, Leibniz Institute for the German Language (IDS), R5, 6-13, 68161, Mannheim, Germany, E-mail: wolfer@ids-mannheim.de. https://orcid.org/0000-0002-8893-8153
Alexander Koplenig, Leibniz Institute for the German Language (IDS), R5, 6-13, 68161, Mannheim, Germany. https://orcid.org/0000-0002-9630-9680

When we compare the frequency of occurrence of several elements over several (sub-)corpora, the absolute frequency of occurrence might not be enough to state that "the element observed more often is more frequent because the observed frequencies are of course dependent on the sizes of the corpus parts that are compared" (Gries 2010: 271). In this case, we have to compute normalised (sometimes also called relative) frequencies. These are typically reported as occurrences in one thousand or one million tokens by multiplying the quotient between raw (or absolute) frequency f_{raw} and corpus size s (number of tokens) by the normalisation quantity c: $f_{\text{norm}} = f_{\text{raw}}/s*c$. Just like the compilation of lists of raw frequencies, the normalisation of frequencies can be considered standard corpus linguistic practice: "frequencies must be normalised before the lists can be compared directly" (Baron et al. 2009).

It is therefore not surprising that not only normalised frequencies per se are omnipresent in corpus linguistic research,[1] but that the *comparison* of frequency values, often of the same linguistic element in (sub-)corpora of differing size, is relevant in many studies. Gries (2010) uses this is as an introductory example when he compares the normalised frequencies of *give* and *bring* across the spoken and written parts of the International Corpus of English (British component, ICE-GB). One of the most prominent examples is what Michel et al. (2011) call "usage frequency" when they normalise the absolute number of *n*-grams by the total number of words in the Google Books corpus in that year. They go on to compare these usage frequencies for single words (e.g., *slavery*) over the years as well as comparing usage frequencies of alternative expressions (e.g., *the Great War*, *World War I*) diachronically. Other comparisons of normalised frequencies between (sub-)corpora of different sizes include, for example, stylometric analyses for two columnists of the *Daily Telegraph* (Grieve 2023) and the study of changes in the use of the epistemic stance markers of modal verbs in climate science over time (Poole and Hayes 2023). In a discourse study of media language in the Egyptian revolution of 2011, Attia and Romero-Trillo (2022) compare, i.a., the frequency of specific keywords (and their collocations) in three corpora of different size (Arabic and English versions of Al Jazeera and Al Arabiya as well as BBC and CNN). Probabilities of non-lexical items are also being compared by calculating normalised frequencies, e.g. laughter in a hospital setting (Macqueen et al. 2024). Here, the sub-corpora are defined by the participant's role in the interactions (patient, nurse, researcher, doctor etc.), and because the different roles contribute differently to the overall corpus, the role-based sub-corpora also differ in size. The great diversity of these examples already suggests how widespread comparisons of normalised frequencies are. The logic is always the

1 A search with the terms "normali[s/z]ed/relative frequenc[y/ies]" in the journal Corpus Linguistics and Linguistic Theory alone results, at the time of writing, in hits for 97 articles.

same: the probability of occurrence of the same or different elements is compared across sub-corpora of different sizes.

However, it is also a well-known fact in corpus linguistics that most, if not all, quantities vary systematically with corpus size. Measures that have been proposed to describe lexical richness of texts "change systematically with the text length" (Tweedie and Baayen 1998: 334). There are also some interesting differences between theoretical constancy, i.e. their mathematical properties given the assumption "that words are used randomly and independently in texts" (Tweedie and Baayen 1998: 332) and empirical constancy, i.e. their actual behaviour in coherent prose. Also, there is an effect of corpus size on the efficiency of a frequency threshold when extracting lexical bundles from corpora when these thresholds "are expressed in normalised frequency" (Bestgen 2018: 205). Furthermore, text similarity measures based on generalised entropies "depend heavily on the sample size" (Koplenig et al. 2019: 1). In the field of psycholinguistics, Burgess and Livesay (1998) show that the predictive power of word frequency measures grows with the corpus size when predicting reaction times for low- and medium-frequency words in a word recognition study.

These are just a few examples where the size of a corpus can have a decisive influence on quantitative measures that are meant to describe (some property of) a corpus or the linguistic material it contains. We therefore wondered whether corpus size might have an even more fundamental influence, namely on normalised frequencies themselves. In this article, we will not specifically evaluate the mathematical properties of normalised frequencies. Rather, we will use corpus data to evaluate the empirical effect(s) of corpus size on normalised frequencies. The basic approach we will use is the pairwise comparison of frequency lists. Each of the two lists originates from a different corpus sample, possibly with a different size (measured in number of included sentences). Overall, we are trying to create a situation that could be considered "optimal" for the calculation and comparison of normalised frequencies, both in terms of the corpus basis (as far as we were able to) and the sampling process. Among other things, we exclude any influences from the text/document level (but not from the sentence level) by using a corpus with scrambled sentences. This also means that we deliberately ignore the internal structure of corpora. As a result, we largely exclude phenomena such as "clumpiness" or "burstiness" (Altmann et al. 2009; Kilgarriff 2001: 241) from the analyses presented here. In addition, we keep the text type as constant as possible by only including generic web corpora for the languages under investigation. In this way, we can be sure that any effects we might find can be attributed solely to corpus size.

We base our analyses on six comparison measures and corpora for six typologically diverse languages and show that corpus size (measured as the number of included sentences) does not have a negative influence on the comparison of

normalised frequencies. Rather, we can show that whenever a larger sample is involved in the comparisons, the association between the lists increases. We also show, though, that the associations systematically decrease for less frequent wordform types.

The remainder of this paper is structured as follows. In Section 2, we will introduce some basic concepts of statistical theory regarding our research question and will formulate two predictions. In Section 3.1, we describe the corpora under investigation and how we pre-processed the data. Section 3.2 describes the sampling process and how we arrived at the sets of wordform types we used for the frequency lists. In Section 3.3, we present the comparison measures we use for measuring the association between frequency lists and how we set up the comparison regimes for lower-frequency wordform types. In the following sections, we first present the results for full (Section 4.1) and truncated (Section 4.2) frequency lists. We discuss the results in Section 5 before wrapping up with the implications of our results (Section 6).

2 Statistical theory

As indicated above,[2] we are trying to create a situation that is "optimal" for calculating and comparing normalised frequencies. Using random samples of sentences is consistent with the random sampling assumption from inferential statistics, at least for the text type (generic web corpora) we have used in this study. Given this assumption, it is further assumed that normalised frequencies are unbiased estimates of the occurrence probability of type i, written as $\widehat{p}_i$.

In the current study, we manipulate sample size s directly via drawing varying amounts of sentences from the base corpora (see Section 3.2). Assuming a binomial sampling distribution, the standard deviation of the expected probability of occurrence of type i is given by the square root of the sampling variance.

$$\sigma_{\widehat{p_i}} = \sqrt{\widehat{p}_i(1-\widehat{p}_i)\Big/ s} \tag{1}$$

Hence, we expect the sampling variation to decrease when the sample size increases. In other words, bigger samples should yield more precise estimates of the probability of occurrence of a specific wordform type.

Furthermore, we can calculate the relative sampling variation in analogy to the coefficient of variation (CV), i.e. the sampling variance relative to p_i. For this, we

2 We would like to thank an anonymous reviewer for pointing out the connections between our empirical results and statistical theory in a very detailed review. This section is strongly inspired by their review.

divide the standard deviation given in eq. (1) by the expected probability of occurrence:

$$\mathrm{CV}_i = \sigma_{\widehat{p_i}} \Big/ \widehat{p}_i \tag{2}$$

CV_i increases when the expected probability of occurrence decreases and vice versa. We must therefore expect that normalised frequency values for low-frequency wordform types are more "unstable" than for high-frequency types. We therefore arrive at two predictions from statistical theory.

1. The more sentences we sample from the base corpora, the more precise the normalised frequency values should be, as indicated by Equation (1).
2. As normalised frequencies decrease, precision should also decrease. Or, in other words, normalised frequency values for low-frequency types are less stable than for high-frequency types, as indicated by Equation (2).

The remainder of the paper can also be seen as an empirical test of these theoretical predictions.

3 Methods

3.1 Data

We use language data from the Leipzig Corpora Collection (Goldhahn et al. 2012). The largest corpora freely available for download on the website[3] contain one million sentences. As we explain in Section 3.2, we need larger corpora for our analyses. We therefore contacted the researchers at the Wortschatz Leipzig (WSL) project directly, who provided us with corpora containing ten million sentences per language.[4] For Chinese (ISO code zho[5]), English (eng), Finnish (fin), French (fra), German (deu), and

3 https://wortschatz.uni-leipzig.de/en/download [last access on 23 May 2025].

4 The corpora containing ten million sentences per language that we base our analyses on are available upon request from the WSL team. The derived frequency lists we use in this paper are available via OSF: https://osf.io/64mpz/ (folder 'Datasets') along with the sets of overlapping word-form types (see Section 2.2), R scripts (R Core Team 2024) for the analyses (folder 'R scripts') and the resulting comparison dataframes (see Section 2.3). On https://wortschatz.uni-leipzig.de/de/download [last access on 23 May 2025], corpora for different languages, years and sources are available including up to one million sentences.

5 Please note that this ISO code indicates the macrolanguage Chinese covering several individual languages with Mandarin Chinese (ISO code cmn) being one of them (for an overview, see https://iso639-3.sil.org/code/zho [last access on 29 October 2024]). The corpus data we received from the WSL team thus includes language material from several of these individual languages.

Vietnamese (vie), ten million randomly shuffled sentences were extracted from generic web corpora of the respective language. In selecting the languages, we were guided on the one hand by the availability of large corpora in the WSL project. English is still the predominant language in corpus linguistics and therefore seemed like the natural choice. While English is fundamentally a Germanic language, much of its vocabulary stems from the Romance languages. Therefore, we chose German and French as additional languages. Chinese and Finnish are seen as two extremes on the analytic-agglutinative continuum. To represent another language family, we chose Vietnamese.

We tokenised and part-of-speech-tagged each corpus with UDPipe, using the R (R Core Team 2024) package {udpipe} (Wijffels 2022) with ready-made models for each of the languages. We excluded all wordforms that span several tokens in the UDPipe tokenisation because the tokenisation process adds the non-contracted elements, e.g., "zu" (Engl. *to*) and "dem" (Engl. *the*) are being added for German "zum" and we exclude the "zum". In French, for example "à" (Engl. *at/in/to*) and "le" (Engl. *the*) are being added for "au" and we exclude the "au".[6] Since the number of sentences is kept constant and sentences differ in length, the number of overall tokens (including punctuation) in the base corpora differs between languages (Chinese: 369,434,986 tokens; English: 227,654,199 tokens; Finnish: 133,665,750 tokens; French: 231,398,746 tokens; German: 182,120,095 tokens; Vietnamese: 211,459,440 tokens).

3.2 Sampling process

For each language, we sampled 100,000 (henceforth "100k"), 500,000 ("500k"), and one million sentences ("1M"). Each size was sampled five times. Whenever a sentence was sampled from the overall corpus of ten million sentences, we excluded this sentence from further sampling. Hence, all 15 samples (five 100k samples, five 500k samples and five 1M samples) are mutually exclusive, i.e. none of the sentences appears twice in the final datasets. This is also the reason why we needed ten million sentences for each language (technically, an overall corpus of $5 \times 100{,}000 + 5 \times 500{,}000 + 5 \times 1{,}000{,}000 = 8{,}000{,}000$ sentences would have sufficed but we wanted to allow for some of the sentences not to be sampled at all).

6 One might wonder why we kept the added tokens and not the original token. It is, unfortunately, much easier to keep the added elements, because UDPipe assigns running token indices to the non-contracted elements (e.g., "zu": 3; "dem": 4) and both indices to the original token ("zum": 3–4). So, we deleted all tokens whose indices included the dash (the original contracted element).

Since we are mainly interested in whether the normalised frequency of a given word form type is affected by corpus size, we need to make sure that this type is actually present in all comparison corpora (= samples and sizes). So, for each language, we computed the set of wordform types that appeared in all 15 samples (henceforth "overlapping types"). The numbers of overlapping types are as follows for each of the six languages: Chinese: 5,086 types; English: 35,699 types; Finnish: 52,345 types; French: 39,410 types; German: 40,266 types; Vietnamese: 18,910 types.[7] This procedure implicates that all overlapping types must have at least an absolute frequency of 15 in the overall ten million sentence corpus.

3.3 Analyses

For each sample, frequencies were normalised to occurrences in one million words. After[8] normalisation, each of the frequency lists was restricted to the overlapping wordform types because we can only compare frequencies of wordform types that occur in all samples for this language.

To answer our main question, whether corpus size has an influence on normalised frequencies, we set up a comparison regime. The logic here is that two frequency lists based on two different corpora should show a high association of normalised frequency values. This is especially true for the corpus samples we use here because the underlying corpus from which we draw is homogeneous (constant text type) and the sentences in the corpus are shuffled. So, if the normalisation of frequencies really works as intended, a differing sample size should not result in lower associations. In fact, we would expect larger samples to work 'better' because the normalised frequency values are based on more data.

---

7 One might wonder why Finnish has more overlapping types than Chinese or Vietnamese. As one anonymous reviewer wrote, "one could expect its [= Finnish] rich morphology to result in a larger pool of possible inflected forms, which wouldn't necessar[il]y occur similarly in the various sample sizes". In fact, the higher number of overlapping types in Finnish is due to the considerably higher number of types in Finnish per se. For example, for 1M sentences, the unrestricted vocabulary size for Finnish is 1.52 million while it is 0.93 million for Vietnamese (averaged over the five 1M samples per language). This higher number of wordform types means that more types 'survive' the selection criterion, i.e. the necessity to appear in all corpus samples.

8 Note that the values for normalised frequencies would differ if we would have restricted the frequency lists to the overlapping types first and then computed the normalised frequencies afterwards. However, this sequence would undermine the purpose of our study because the normalisation of frequencies must consider the total token frequency for the entire sample.

We compared each sample frequency list to all the other sample frequency lists for the same language. We used the following measures for these comparisons:[9]

- **Kendall's rank correlation coefficient** τ_B, which measures the ordinal association between the two frequency lists by comparing the number of concordant and discordant pairs of observations while accounting for tied pairs. A pair is concordant if both differences between two data points in a bivariate series of measurements point in the same direction. Accordingly, a pair is discordant if the differences point in the opposite direction.[10]
  We used the considerably faster (compared to cor(…, method = "kendall") in {stats}) implementation in the {pcaPP} R package (Filzmoser et al. 2023). This algorithm goes back to Knight (1966) and is being described in more detail by Abrevaya (1999) and Christensen (2005). Higher values of τ_B indicate a stronger association between two frequency lists and it ranges between –1 (perfect negative association) and 1 (perfect positive association).
- The **proportion of concordant pairs** p_C between the two frequency lists, which is one component of the calculation of τ_B. Higher values of p_C indicate a stronger association between the two frequency lists. p_C ranges between 0 and 1. p_C is defined as

$$p_C = \frac{n_C}{n_C + n_T + n_D} \tag{3}$$

  where n_C is the number of concordant, n_T the number of tied and n_D the number of discordant pairs.
- The **proportion of concordant and tied pairs** p_{CT} between the two frequency lists. We included this measure because tied pairs can also be indicative of an

9 Note that, in principle, it should not matter if we compare raw or normalised frequencies for all comparison measures that only involve frequency ranks. However, in our scenario, normalising frequencies might yield slightly different results from those obtained with raw frequencies because the (sub-)corpus size (= number of tokens) differs between samples, even when the same number of sentences is sampled. So, a wordform type with the same raw frequency in both frequency lists might get assigned different normalised frequency values which in turn impacts the calculation of said comparison measures. However, τ_B, p_C, and p_{CT} for raw versus normalised frequencies all correlate strongly over all frequency list comparisons (smallest Spearman's r = 0.996 for p_{CT}). We will report the variants calculated on normalised frequencies.

10 For illustrational purposes, consider a series of bivariate measurements with x = [3, 2, 1, 1] and y = [5, 7, 6, 5]. The first data points versus the second data points (3 vs. 2 and 5 vs. 7) yields a discordant pair, the first versus the third (3 vs. 1 and 5 vs. 6), too. The first versus the fourth (3 vs. 1 and 5 vs. 5) counts as tied because of the tie in y. The second versus the third (2 vs. 1 and 7 vs. 6) is concordant, just like the second versus the fourth (2 vs. 1 and 7 vs. 5). The final pair (1 vs. 1 and 6 vs. 5) is tied because of the tie in x. Hence, we arrive at 2 concordant pairs, 2 discordant pairs, and 2 tied pairs.

association between the two lists and together with p_C, we get a better insight into the procedure for calculating τ_B. Also, by comparing p_C and p_{CT}, we are able to track down potential effects of the number of tied pairs alone. Higher values of p_{CT} indicate a stronger association and its maximum value is 1. Note that, by definition, p_{CT} is always higher or equal to p_C because it is defined as

$$p_{CT} = \frac{n_C + n_T}{n_C + n_T + n_D} \tag{4}$$

– The mean normed **deviation of normalised frequencies** Δ_f between the two frequency lists A and B. This measure is defined as

$$\Delta_f = \frac{\sum_{i=1}^{N} |a_i - b_i| \,/\, m_i}{N} \tag{5}$$

where N is the number of wordform types on the two frequency lists, a_i and b_i are the normalised frequencies of type i on list A and list B, respectively, and m_i is the arithmetic mean of a_i and b_i. Note that the absolute deviation of normalised frequencies (without the norming by m_i) would also be possible. However, this would underestimate the deviations for low-frequency types because the "potential" for deviations is much higher for higher-frequency types.[11] The optimal value of Δ_f is 0 which would indicate a perfect match of the normalised frequencies of all types.

We have included this comparison measure because, unlike τ_B, p_C and p_{CT}, it does not operate on the basis of pairs but implements the comparison of two frequency values for the same word form type in a more direct way.

– The **proportion of types with a higher normalised frequency on list A than on list B** p_{HH} (HH stands for "half-half"). The rationale behind this comparison measure is that, if normed frequencies, indeed, fluctuate randomly between two frequency lists, around 50 % of all wordform types on list A should have a higher normalised frequency than on list B (the same is the case for the other direction). p_{HH} is defined as

11 To illustrate this point, say we have two samples of identical size (in tokens). Consider a type t_1 with a frequency of 1 on list A and a frequency of 3 on list B. The absolute difference of normalised frequencies would also be very small. In contrast, consider a type t_2 with a frequency of 10,000 on list A and a frequency of 10,050 on list B. Here, the absolute difference of normalised frequencies would be much higher than for t_1 although the two frequencies might be considered "more equal". The numerically higher deviation is only because t_2 is located in a higher frequency band than t_1. This can be corrected by normalising the absolute difference with the mean value of the two normalised frequencies.

$$p_{\mathrm{HH}} = \frac{N_{A>B}}{N} \tag{6}$$

where $N_{A>B}$ is the number of wordform types with a higher normalised frequency on list A than on list B. The optimal value of p_{HH} is 0.5 which would indicate that exactly half of the types have a higher normalised frequency value on list A than on list B (and the other way around). There are no cases in our datasets where the normalised frequencies of any wordform type are exactly equal.

– The **mean normalised frequency ratio** r_f (cf. Gries 2010: 272) between the normalised frequencies of any given wordform type on list A and B. r_f is defined as

$$r_f = \frac{\sum_{i=1}^{N} a_i \Big/ b_i}{N}. \tag{7}$$

The optimal value for r_f is 1. Note, however, that a value of 1 does not necessarily mean that all ratios are exactly 1 (i.e. that all pairs of a_i and b_i are equal) but that the *mean* of these ratios is 1.

To investigate whether the frequency range of the wordform types (from highly frequent to very infrequent types) has an impact on the comparison measures, we subsequently exclude types from the top of the frequency lists. We used the following steps: i) all types are being used, ii) 80 % of types from the bottom of the frequency list are being used, i.e. 20 % of the top frequency types are being excluded, iii) bottom 60 %, iv) bottom 40 %, v) bottom 20 %, i.e. 80 % of the top frequency types are being excluded. Results for these truncated frequency lists are presented in Section 4.2.

All in all, there are 3 (sizes of sample 1) × 5 (sample 1 iterations) × 3 (sizes of sample 2) × 5 (sample 2 iterations) × 5 (exclusions) = 1,125 comparisons for each language. From these, 3 (sizes) × 5 (samples) × 5 (exclusions) = 75 comparisons compare the frequency list with itself, so we conduct 1,125 − 75 = 1,050 comparisons per language. For each of these comparisons, we compute the six measures outlined above.

4 Results

4.1 Full frequency lists

Figure 1 presents the results for Kendall's rank correlation coefficient (*y*-axis) for each language (colour and shape) and all combinations between the size of the first sample (*x*-axis) and the size of the second sample (plot panel). Each data point stands for one

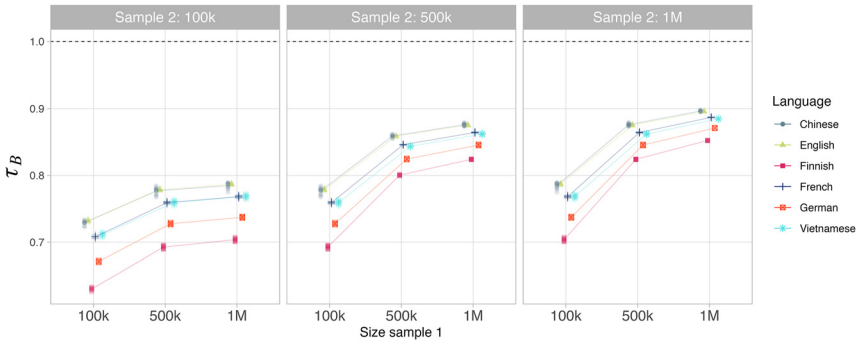

Figure 1: The influence of the size of sample 1 (*x*-axis) and sample 2 (plot panels) on Kendall's τ_B (*y*-axis) measuring the correspondence of ranks for all types on the two frequency lists. Each data point is one comparison. The mean values for each group of data points are connected by lines. Languages are distinguished by colour and point symbol. The maximum value of $\tau_B = 1$ is marked by the dashed line. Note: The lines are only inserted to make the distinction between languages easier, i.e. there are no data points between the tick marks on the *x*-axis. The languages are slightly dodged on the *x*-axis to allow for easier distinction, i.e. sample sizes are all 100k, 500k or 1M sentences.

comparison (20 per sample-sample combination) and the lines connect the mean values of each group of data points.[12] The dashed line indicates the optimal value of $\tau_B = 1$.

The different comparison data points within each group are not dispersed over a wide range of τ_B values indicating little dispersion between samples. For example, in the left-most yellow/triangle group (all comparisons for the English corpus comparing 100k sentence samples with all other 100k sentence samples), the values of τ_B vary between 0.729 and 0.734. Also, there is no systematic difference of effect patterns between languages. The central question, however, is whether sample size has an effect on τ_B. That is, indeed, the case. Whenever a larger sample is involved in the comparison, τ_B increases. Consequently, the smallest value of τ_B is obtained for 100k versus 100k comparisons and the largest value is obtained for 1M versus 1M comparisons. These maximal observed values (mean τ_B for Chinese: 0.896, English: 0.896, Finnish: 0.852, French: 0.887, German: 0.871, Vietnamese: 0.885) are close to the theoretical maximum of 1.

Given this pattern for comparisons via τ_B, we conclude that corpus size indeed has an effect on the comparability of normalised frequencies. But it is exactly the kind of influence we would expect from larger corpus samples given prediction 1 from Section 2. Even though the inclusion of a larger sample leads to very disparate sample sizes

12 Please note that some data points are redundant because comparisons are symmetric. For example, the data points for 100k versus 500k are identical to the ones for 500k versus 100k. We kept these comparisons anyhow to allow for easier comparisons within each plot panel.

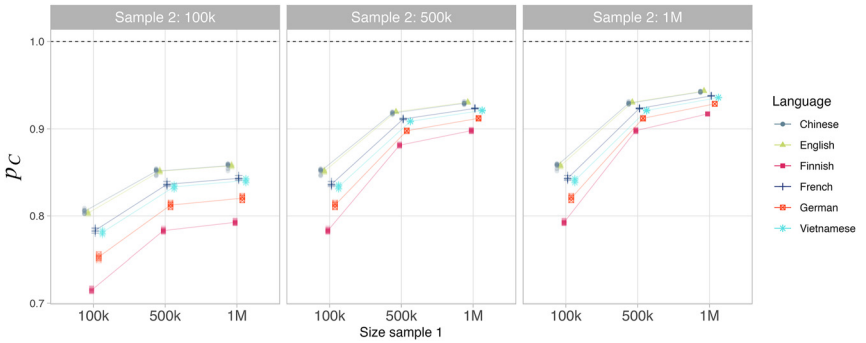

Figure 2: The influence of the size of sample 1 (*x*-axis) and sample 2 (plot panels) on the proportion of concordant pairs p_C (*y*-axis). Plot organisation is equivalent to Figure 1.

(in our case, the most extreme difference in sample size is 100k vs. 1M sentences), τ_B systematically increases, and this holds for all languages under investigation.

Figures 2 and 3 basically show the same pattern as Figure 1: as soon as (at least) one of the two corpus samples in the comparison is larger, p_C and p_{CT} increase with maximum mean values of 0.942 (p_C, Chinese), 0.944 (p_C, English), 0.917 (p_C, Finnish), 0.938 (p_C, French), 0.929 (p_C, German), 0.936 (p_C, Vietnamese) and 0.950 (p_{CT}, Chinese), 0.950 (p_{CT}, English), 0.930 (p_{CT}, Finnish), 0.946 (p_{CT}, French), 0.938 (p_{CT}, German), 0.945 (p_{CT}, Vietnamese).

For Δ_f, we observe effect patterns similar to the previous comparative measures (see Figure 4). This time, however, a better match between the frequency lists is indicated by a lower value (with the theoretical minimum being 0). Again, as soon as a frequency list based on a larger corpus sample is part of the comparison, the measure is closer to the optimum value. Since Δ_f is the mean normed deviation between the normalised frequencies of each type, this result could be influenced by a few outlier types with very high absolute deviations. However, if we calculate the *median* normed deviation (see Supplementary Figure 1), this effect pattern does not change considerably. Overall, the median values are slightly lower (e.g., for English, the range for the mean normed deviation is 0.0247–0.0779, for the median normed deviation, it is 0.0132–0.0341). These findings also generalise over languages.

For p_{HH}, we only plot symmetric comparisons, i.e. where samples 1 and 2 contain the same number of sentences. Thus, we can lose the distinction via plot panels in Figure 5.[13] With larger sample sizes, the data points for the different

13 We only compare equally sized samples here because of the directedness of the comparison measure: p_{HH} is defined as the proportion of types that are more frequent on frequency list A than on list B. So, if we would compare, for example, a 100k (list A) with a 500k frequency list (list B), we know that the data points must have the same distance to 0.5 than for the other direction of this comparison

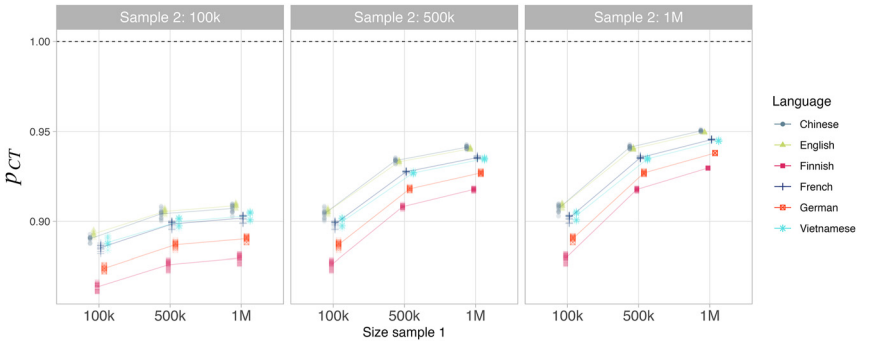

Figure 3: The influence of the size of sample 1 (*x*-axis) and sample 2 (plot panels) on the proportion of concordant and tied pairs p_{CT} (*y*-axis). Plot organisation is equivalent to Figure 1.

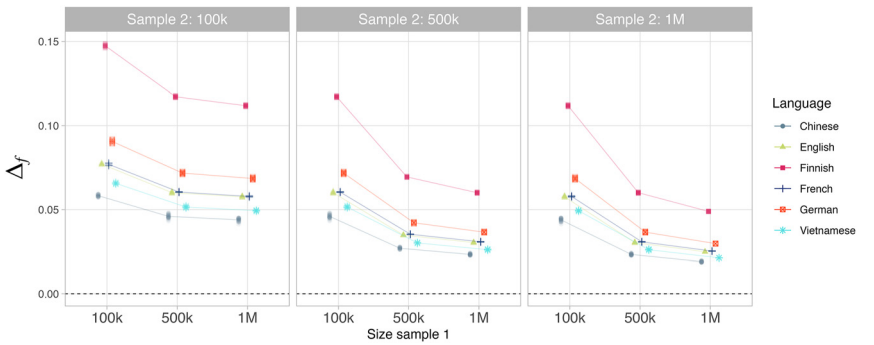

Figure 4: The influence of the size of sample 1 (*x*-axis) and sample 2 (plot panels) on the mean normed deviation of normalised frequencies Δ_f (*y*-axis). Plot organisation is equivalent to Figure 1.

symmetric comparisons lie closer to the optimum value. Hence, the maximum distance to $p_{HH} = 0.5$ is for a 100k versus 100k comparison (e.g., for French: 0.425 and 0.575 for the symmetric comparison). The minimum distance can be observed for a 1M versus 1M comparison (again, for French, 0.477 and 0.523). The comparisons for 500k versus 500k sentence samples lie in between, and this holds for all languages.

For the mean normalised frequency ratio (see Figure 6), we see that, again, as soon as one larger sample is involved in the comparison, the comparison measure

(100k: list B; 500k: list A). However, since the different sizes of sample 2 (= frequency list B) are distributed over plot panels, these symmetric comparisons would also be distributed over different panels which would make for a rather confusing visualization. We provide the annotated Supplementary Figure 2 to illustrate this.

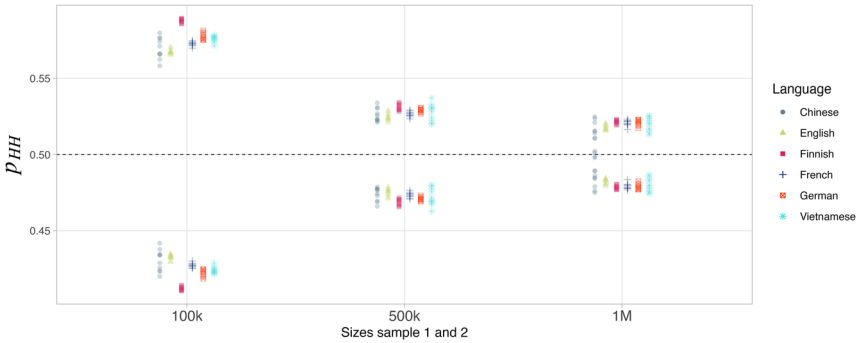

Figure 5: The influence of the sizes of samples 1 and 2 (*x*-axis) on the proportion of word types that is more frequent in sample 1 than in sample 2 p_{HH} (*y*-axis). Each data point is one comparison. Languages are distinguished by colour and point symbol.

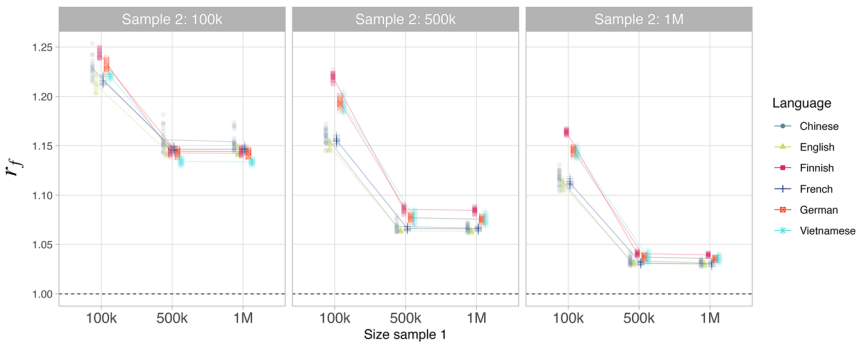

Figure 6: The influence of the size of sample 1 (*x*-axis) and sample 2 (plot panels) on the mean normalised frequency ratio r_f (*y*-axis). Plot organisation is equivalent to Figure 1.

approaches its optimum value (here $r_f = 1$). For example, for German, we observe the lowest mean value for the 1M versus 1M comparison (mean $r_f = 1.035$). The mean value for the 500k versus 1M comparison is very close though (mean $r_f = 1.037$).

4.2 Truncated frequency lists

We proceed by replicating the analyses from the previous section and then restricting the analyses in a stepwise manner to lower-frequency types. We do this to investigate whether the size of the samples that enter the comparison has a different

effect on types in different frequency ranges. We start with the full frequency lists, i.e. replicating the analyses from the previous section and exclude 20 % more in each step, yielding five steps where the last step only involves the 20 % of types from the bottom of the frequency lists. We report the raw number of included types in each step for all languages in Supplementary Table 1.

Figure 7 (plot A) shows the results for Kendall's τ_B. First, there is a clear and continuous influence of the number of included types, as given in prediction 2 from Section 2: the fewer types we include, the lower the correspondence of ranks as indicated by τ_B. Secondly, we see a similar pattern distinguishing between the sample sizes as in the previous section: as soon as one larger sample is involved in the comparison, correspondence levels increase, this is both visible from the left to the right panels for each language as well as the relative positions of the lines connecting the means for each group of data points. By comparing the panel rows, we see that these effects are consistent over the six languages. Figure 7 (plot B) replicates this 'behaviour' for one component of τ_B, the concordant pairs.

What is interesting, though, is the pattern in Figure 8 (plot A) where the proportion of concordant *and tied* pairs is visualised. While the overall pattern remains largely stable, there are slight deviations for the last step where we only include the bottom 20 % of types. This effect is especially pronounced for the comparisons between the smallest samples (100k vs. 100k sentences) in all investigated languages. Here, we see that p_{CT} indicates a *better* correspondence than in previous steps and in the comparisons of larger samples. This contradicts all previous findings. It is thus interesting and important to track down the cause for this pattern. Since this picture could not be observed for p_C, it must be related to the tied pairs, the number of which is apparently 'irregularly' higher for the smallest comparisons. This is solely due to the number (or proportion) of hapax legomena:[14] while this proportion is less than 2 % for the larger comparisons, an average of 19.2 % of all types can only be observed once in both samples for the smallest comparisons for English (30.5 % for Finnish, 20.7 % for French, 25.7 % for German, 26.4 % for Vietnamese).[15] All hapax legomena 'produce' tied pairs, therefore increasing p_{CT}. Similar distortions can be observed for the mean normed deviation of normalised frequencies (Figure 8, plot B) and the mean relative frequency ratio (Figure 9, plot B).

14 Please note that the status as hapax legomenon refers exclusively to the samples and not to the underlying corpora we sampled from. Each overlapping wordform type had to occur at least 15 times in the respective underlying corpus, but not necessarily more than once in each sample. See Section 3.2 for details.

15 The exact mean proportions of hapax legomena for the bottom 20 % of types are given in Supplementary Table 2 for all languages. The highest mean proportion for comparisons of larger samples is 1.74 % for 100k versus 500k samples for Finnish.

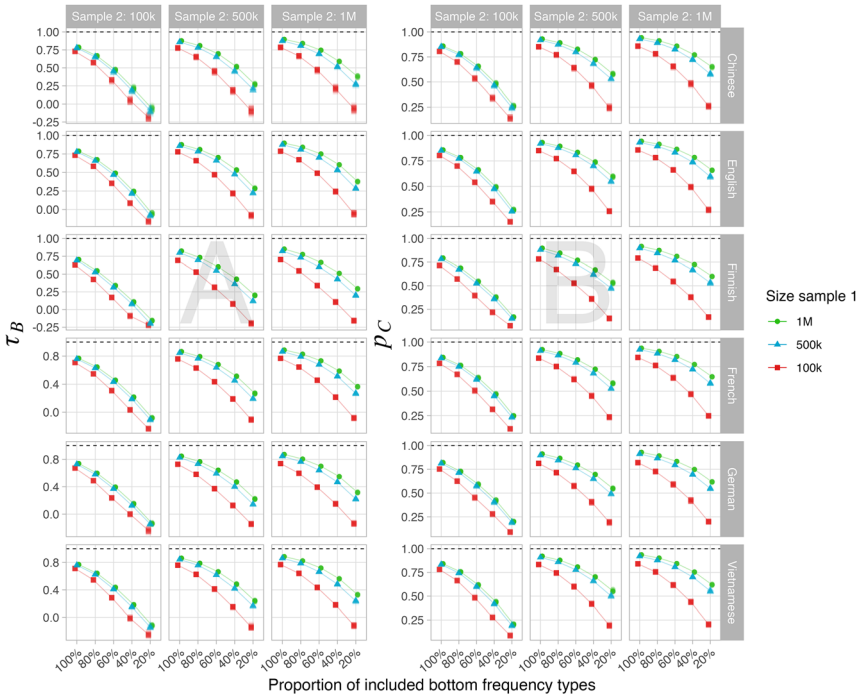

Figure 7: The influence of the number of included types from the bottom of the frequency lists (*x*-axis), size of sample 1 (point colour and symbol) and sample 2 (plot panels, column-wise) on Kendall's τ_B (*y*-axis, plot A) and the proportion of concordant pairs p_C (plot B). The lines connect the mean values for each group of data points. The maximum values of $\tau_B = 1$ and $p_C = 1$ are marked with dashed lines. Languages are distinguished by row-wise plot panels.

For Δ_f (see Figure 8, plot B), the optimum value is 0, hence lower values indicate higher correspondence between frequency lists. Apart from the effect of hapax legomena on the smallest comparisons (20 % of bottom frequency types, 100k vs. 100k sentence samples), its trajectory indicates higher correspondences for larger samples and lower association when fewer types are included in the comparisons. Again, this holds for all languages under investigation.

For p_{HH} (see Figure 9, plot A), we again only plot symmetric comparisons, i.e. the comparisons between equally-sized samples. The development over the number of included types follows a funnel-shaped pattern: The more we restrict the analyses to rarer types, the further the values move away from the optimum value of $p_{\mathrm{HH}} = 0.5$. Again, the values for larger samples are closer to the optimum

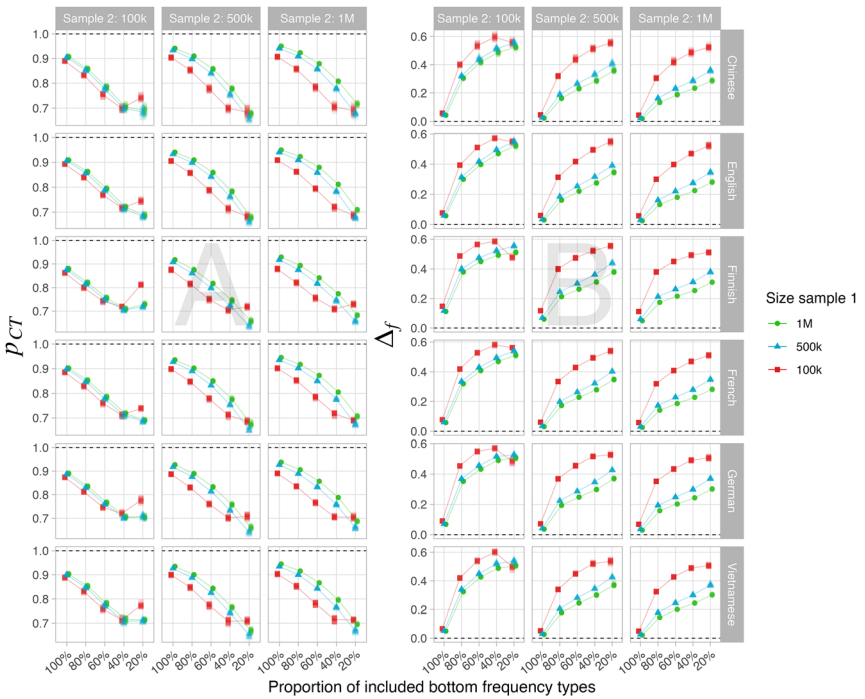

Figure 8: The influence of the number of included types from the bottom of the frequency lists (*x*-axis), size of sample 1 (point colour and symbol) and sample 2 (plot panels, column-wise) on the proportion of concordant and tied pairs p_{CT} (*y*-axis, plot A) and the mean normed deviation of normalised frequencies Δ_f (plot B). The lines connect the mean values for each group of data points. The maximum values of $p_{CT} = 1$ and $\Delta_f = 0$ are marked with dashed lines. Languages are distinguished by row-wise plot panels.

value. The only data point that does not follow this pattern is the smallest comparison for French (fourth panel row, 100k vs. 100k, 20 % included types) which is slightly *closer* to 0.5 than the data point for twice as many included types. We currently have no explanation for this effect, especially why it only holds for one of the languages under investigation.

The distortions for small sample comparisons (bottom 20 % of types, 100k vs. 100k) due to hapax legomena are quite pronounced for the mean relative frequency ratio r_f (Figure 9, plot B). In Figure 6, we already saw that the differences for r_f between 500k and 1M comparisons were rather small. This does not only hold for the full frequency lists. For each language, the lines for 500k and 1M converge if 60 % or more of the bottom frequency types are included.

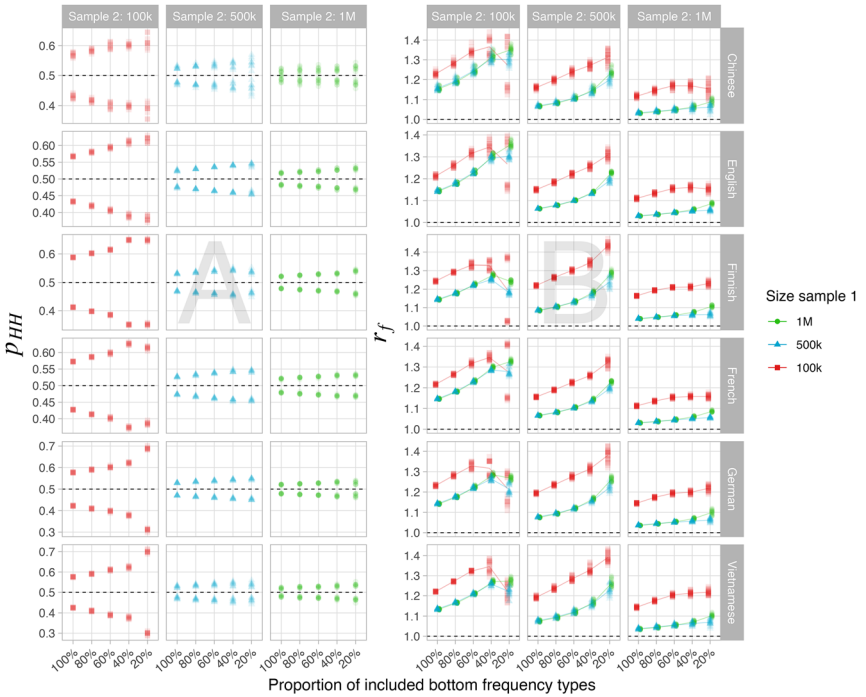

Figure 9: The influence of the number of included types from the bottom of the frequency lists (*x*-axis), size of sample 1 (point colour and symbol) and sample 2 (plot panels, column-wise) on proportion of word types that is more frequent in sample 1 than in sample 2 p_{HH} (*y*-axis, plot A) and the mean normalised frequency ratio r_f (*y*-axis, plot B). The optimal values of $p_{HH} = 0.5$ and $r_f = 1$ are marked with dashed lines. Languages are distinguished by row-wise plot panels.

5 Discussion

Several frequency-based corpus linguistic measures are strongly influenced by corpus size (e.g. measures of lexical diversity or text similarity metrics). Based on this observation, we empirically investigated whether the very basal measure of normalised frequencies is also influenced by corpus size. To our knowledge, it is largely unquestioned that normalised frequencies are supposed to correct for the influence of corpus size – but it has not yet been systematically tested empirically how they might be influenced by corpus size themselves. We approached this question by comparing frequency lists, or rather testing the association that exists between two lists of normalised frequencies. For generic web corpora of Chinese, English, Finnish, French, German, and Vietnamese, we kept all influencing factors as constant as

possible and only systematically varied the size of the underlying corpus from which the frequency lists were derived.

The analyses for the complete frequency lists paint a consistent picture and confirm predictions from statistical theory, both over languages and comparison measures: whenever a larger corpus sample is involved in the comparison, the six comparison measures for the frequency lists derived from those samples are closer to the respective optimal value (e.g. 100k vs. 500k or 500k vs. 1M sentences). So, it is *not* the case that the association is always higher when the sample sizes are the same (e.g. 100k vs. 100k or 500k vs. 500k sentences) – an assumption that would logically follow if one would assume that differences in corpus size would systematically skew normalised frequencies.

At the same time, however, we see that the association strengths indicated by the measures decrease rapidly when we restrict the comparisons to lexical items that lie in lower frequency bands, again confirming the prediction based on statistical theory. This means that the comparison of normalised frequencies for these elements is less reliable. On the other hand, the main finding also applies here: as soon as a larger sample is involved in the comparisons, we observe higher associations of the frequency lists, just at a lower overall level. This is what one would expect based on the Zipfian distribution of language data (Zipf 1935), because the discriminatory power of frequencies decreases in lower frequency ranges.[16] An example with two low-frequency word form types: in the five English samples with 100k sentences, the type "glide" is more frequent than "stitch" in two samples, equally frequent in one sample and less frequent in two samples (an overview of raw and normalised frequencies for the example types is given in Supplementary Table 3). In contrast, the relationship is consistent for two high-frequency types: "on" is more frequent than "with" in every sample. Now, if we consider the largest samples in our data set, which are based on 1M sentences each, the ranking between "glide" and "stitch" is stable: "glide" is consistently more frequent than "stitch". In this sense, larger corpora "help" because they are more sensitive to frequency differences in (previously) lower frequency ranges. However, it should be noted that these larger corpora have the same problem in their own lower frequency ranges. In other words, the problem "shifts" to even rarer words.

16 Müller-Spitzer et al. (2015: 13f) make a similar point for predicting look-up frequencies of online dictionary articles by the corpus frequencies of the entries' headwords. There is a clear positive effect for the first few tens of thousands of words at the top of the frequency list. As corpus frequency decreases, though, the effect on look-up frequency gets weaker. The authors claim that this is not because the effect does not exist for headwords with lower corpus frequencies, but because they cannot measure it. Corpora get less and less sensitive to frequency differences in lower frequency ranges.

Regarding the comparative measures we have used, τ_B, p_C and p_{CT} are redundant to a certain extent. However, we could see in the contrast between Figures 7 and 8 (plot A) that of these three measures, only p_{CT} was able to draw attention to the high proportion of hapax legomena for the 20 % least frequent of overlapping types, which may give this measure some justification. r_f shows certain weaknesses when comparing larger corpus samples in the sense that it can no longer cleanly distinguish the associations in frequency lists based on 500k sentences versus 1M sentences.

6 Implications

Our main result is reassuring: the corpus size of the underlying corpora does not negatively influence comparisons of normalised frequency lists, i.e. differences in corpus size do not lead to lower associations between the derived lists. Rather, the results suggest that larger corpora (even in combination with smaller corpora) always have a positive effect in the sense that the association of the compared lists increases. However, we can only draw this conclusion for the scenario we have chosen for the present study, i.e. if all influences except the corpus size (e.g., text type[17] or sampling processes) are kept as constant as possible. For smaller corpora in particular, however, the comparability for less frequent types can quickly decrease. It would be interesting to see whether these findings also hold for more heterogeneous base corpora, for example of varying modality or text type. As already mentioned in the Introduction, we wanted to create a setup here that can be considered "optimal" for the comparison of normalised frequencies.

So, does this mean that larger corpora are a cure-all when it comes to comparing normalised frequencies? Given that other potential sources of influence (e.g., corpus type or sampling processes) were kept as constant as possible, our results suggest that larger corpora always have a positive effect in the sense that the association of the derived frequency lists increases – just like statistical theory predicts. This is also true when larger corpora are combined with smaller ones. On the other hand, especially with smaller corpora, the comparability for less frequent types can quickly decrease. Against this background, our results thus indicate that, all other things being equal – taking into account important considerations such as corpus quality, its

17 It should be noted here that the fact that we only used web corpora does not necessarily imply that the text type is actually constant across languages. For example, as one reviewer quite rightly pointed out, it can be assumed that a lot of material, especially for English web corpora, actually stems from non-native speakers. We cannot completely dispel these concerns, but present for one of our measures (τ_B) in Supplementary Figure 3 that the ranking of languages remains constant when we use newspaper corpora instead of web corpora for English, French, and German.

composition, and the balance among various text types (Biber 1993; Koplenig 2017; Koplenig 2019; Leech 2007) – Mercer's claim that "more data is better data" (Church and Mercer 1993) holds validity.

We would also like to point out that some of the pre-processing steps that we have adopted here cannot equally be applied to all kinds of linguistic research questions. For example, we have excluded very low-frequency words that may be relevant for, e.g., productivity studies. In addition, we have deliberately destroyed the internal structure of the corpora, i.e. the sequence of sentences and their assignment to individual corpus documents, which may be highly relevant for other linguistic research endeavours. It therefore remains to be seen to what extent the results shown here can be transferred to these types of studies.

Acknowledgments: We would like to thank the team of the Leipzig Corpora Collection for providing us with the 10M sentence corpora.
Data availability: The Supplementary Material, R scripts and data used for the analyses in this manuscript are available via OSF at https://osf.io/64mpz/. Please see the Wiki page of the project for explanations concerning the availability of the underlying corpora.

References

Abrevaya, Jason. 1999. Computation of the maximum rank correlation estimator. *Economics Letters* 62(3). 279–285.

Altmann, Eduardo G., Janet B. Pierrehumbert & Adilson E. Motter. 2009. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS One* 4(11). e7678.

Attia, Safa & Jesus Romero-Trillo. 2022. A corpus-assisted discourse study of the media language in the Egyptian revolution. *Language, Discourse & Society* 10(2). 105–127.

Baron, Alistair, Paul Rayson & Dawn Archer. 2009. Word frequency and key word statistics in historical corpus linguistics. *Anglistik* 20(1). 41–67.

Bestgen, Yves. 2018. Evaluating the frequency threshold for selecting lexical bundles by means of an extension of the Fisher's exact test. *Corpora* 13(2). 205–228.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243–257.

Burgess, Curt & Kay Livesay. 1998. The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kučera and Francis. *Behavior Research Methods, Instruments, & Computers* 30(2). 272–277.

Christensen, David. 2005. Fast algorithms for the calculation of Kendall's τ. *Computational Statistics* 20(1). 51–62.

Church, Kenneth W. & Robert L. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics* 19(1). 1–24.

Filzmoser, Peter, Heinrich Fritz & Klaudius Kalcher. 2023. pcaPP: Robust PCA by projection pursuit. Available at: https://CRAN.R-project.org/package=pcaPP.

Goldhahn, Dirk, Thomas Eckart & Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)*, 759–765. Istanbul, Turkey: European Language Resources Association (ELRA). Available at: http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf.

Gries, Stefan Th. 2010. Useful statistics for corpus linguistics. In Aquilino Sánchez & Moisés Almela (eds.), *A mosaic of corpus linguistics: Selected approaches*, 269–291. Frankfurt am Main: Peter Lang.

Grieve, Jack. 2023. Register variation explains stylometric authorship analysis. *Corpus Linguistics and Linguistic Theory* 19(1). 47–77.

Kilgarriff, Adam. 2001. Comparing corpora. *International Journal of Corpus Linguistics* 6(1). 97–133.

Knight, William R. 1966. A computer method for calculating Kendall's Tau with ungrouped data. *Journal of the American Statistical Association* 61(314). 436–439.

Koplenig, Alexander. 2017. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets – Reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities* 32(1). 169–188.

Koplenig, Alexander. 2019. Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory* 15(2). 321–346.

Koplenig, Alexander, Sascha Wolfer & Carolin Müller-Spitzer. 2019. Studying lexical dynamics and language change via generalized entropies: The problem of sample size. *Entropy* 21(5). https://doi.org/10.3390/e21050464.

Laurence, Anthony. 2021. What can corpus software do? In Anne O'Keeffe, Michael McCarthy (eds.), *The Routledge handbook of corpus linguistics*, 2nd edn., 103–125. Abingdon, Oxon; New York, NY: Routledge.

Leech, Geoffrey. 2007. New resources, or just better old ones? The holy grail of representativeness. In Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (eds.), *Corpus linguistics and the web*, 133–149. Amsterdam: Rodopi.

Macqueen, Susy, Luke Collins, Gavin Brookes, Zsófia Demjén, Elena Semino & Diana Slade. 2024. Laughter in hospital emergency departments. *Discourse Studies* 26(3). 358–380.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martina A. Nowak & Erez Lieberman Aiden & The Google Books Team. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014). 176–182.

Müller-Spitzer, Carolin, Sascha Wolfer & Alexander Koplenig. 2015. Observing online dictionary users: Studies using wiktionary log files. *International Journal of Lexicography* 28(1). 1–26.

Poole, Robert & Nicholas Hayes. 2023. Stance in climate science: A diachronic analysis of epistemic stance features in IPCC physical science reports. *Journal of Corpora and Discourse Studies* 5. 37–60.

R Core Team. 2024. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org.

Tweedie, Fiona J. & R. Harald Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32(5). 323–352.

Wijffels, Jan. 2022. udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the "UDPipe" "NLP" toolkit. Available at: https://CRAN.R-project.org/package=udpipe.

Zipf, George Kingsley. 1935. *The psycho-biology of language*. Mifflin: Oxford: Houghton.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/cllt-2024-0040).