Research Methods in Applied Linguistics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Toward more careful corpus statistics: uncertainty estimates for frequencies, dispersions, association measures, and more

Stefan Th. Gries

University of California, Santa Barbara & Justus Liebig University Giessen, Germany

A R T I C L E I N F O

Keywords:
Corpus linguistics
Frequency
Dispersion
Association
Bootstrapping
Interval estimates

A B S T R A C T

This article demonstrates that, counter to current practice, (i) corpus-linguistic studies should provide uncertainty/interval estimates for all corpus-linguistic statistics, even for basic/fundamental ones such as frequencies, dispersions, or association measures, and (ii) these statistics should be based on text-/file-based bootstrapping and confidence/data ellipses covering two or more dimensions of information. Four small case studies – three more programmatic and one more applied – are offered to exemplify the logic and method. The first case study shows how parametric confidence intervals or confidence intervals from word-based bootstrapping can be inappropriate; the second case study exemplifies the computation of frequency-cum-dispersion intervals; the third does the same for collocational/collostructional data (the ditransitive); and the last case study exemplifies the use of these methods in a diachronic statutory-interpretation context.

1. Introduction

Nearly all corpus-linguistic studies at some point report some basic statistical results such as

- the frequency/ies of (co-)occurrence of some element(s) in a (part of a) corpus.
- the dispersion(s) of some element(s) in (parts of) a corpus.
- association measures quantifying (dis)preferences of co-occurrence of two (kinds of) elements in (parts of) a corpus.

As a trivial example, one might state that the word *give* occurs 444 times in the British component of the International Corpus of English (the 1m words ICE-GB), and many studies restrict themselves to reporting observed frequencies (sometimes normalized to per-million-words (pmw)). However, reporting the frequency, while still the standard, is sub-optimal because it neglects the fact that, in nearly all corpus-linguistic studies, the corpus being studied is a sample of a population it is supposed to represent.[1] This means that, as corpus linguists, we need to be constantly aware of sampling variation, but too often we do not seem to be. Many experimental studies in linguistics (Second Language Acquisition (SLA), psycholinguistics, or applied linguistics) report descriptive statistics such as means and standard deviations for relevant experimental groups, but corpus studies – studies in learner corpus research, historical linguistics, descriptive synchronic research, etc. – rarely do so. They normally do *not* provide corpus frequencies together with estimates/indications of uncertainty/variability, namely a measure of dispersion (in either the statistical or the corpus-linguistic sense). There are some exceptions such as text-linguistic or register-analytic work by scholars like Biber and Egbert (see, for instance, Biber & Egbert 2018), but overall, the measures of dispersion in either sense are still too rare.

How might one do this? The most apparent solution is to report the frequency of the form *give* in the ICE-GB as 444 and then report its 95% confidence interval (CI) ([404.6, 487.3], as computed with functions such as prop.test (or binom.test) in R. However, while

E-mail address: stgries@linguistics.ucsb.edu

[1] One might even consider this a defining characteristic of a corpus, one that would distinguish prototypical corpora from databases (e.g., one might prefer to call a collection of all and only all of Shakespeare's literary writings a database rather than a corpus).

https://doi.org/10.1016/j.rmal.2021.100002
Received 19 August 2021; Received in revised form 13 November 2021; Accepted 16 November 2021
Available online xxx

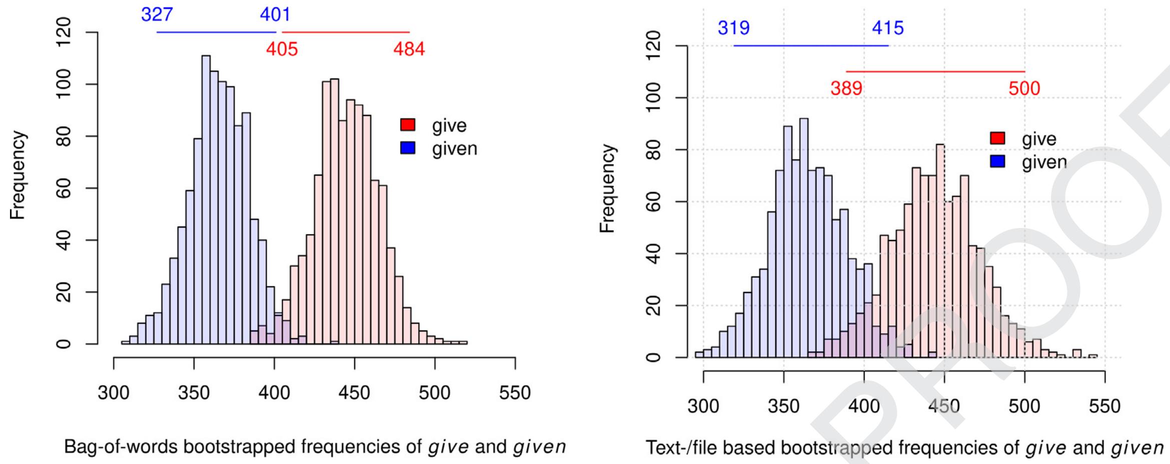

Fig. 1. Comparing confidence intervals for give/given in the ICE-GB.

giving any indication of the statistical dispersion of *give* is unfortunately rare and, thus, would be laudable, the numbers provided are still problematic. To explain how and why, consider the following scenario: You do a corpus-linguistic analysis of something where, for your analysis to be supported by the corpus data, *give* would need to be more frequent than *given*. You look up the frequencies of the two words in the ICE-GB, and you find that *give* and *given* occur 444 and 365 times, respectively. You write it up and send it off for publication. However, some pesky reviewer asks

> But how does the author know that the difference between 444 and 365 is “significant” or “robust enough” (i.e., not just a sampling error or something like that)? I would like to see the authors demonstrate that we can place enough trust in this frequency difference to consider their research hypothesis as supported.

The editor agrees with the reviewer’s comment. Hence, you decide to provide the above kind of confidence intervals for the two frequencies: All it takes are two lines of R code, and you are happy to report that the 95% CIs of *give* and *given* do not overlap, meaning that you have the significant result you hoped for:

- 95% CI of *give*: [404.6, 487.3];
- 95% CI of *given*: [329.4, 404.4].

However, just before you send this off for publication, another review comes in late, which agrees with the idea of requiring you to provide CIs, but recommends a bootstrapping approach, which is an approach that involves repeated sampling with replacement from one’s data in order to estimate how variable the results obtained from the data are (and with how big a grain of salt the results might have to be taken); see Gries (2006) for an early application in corpus linguistics; Egbert and Plonsky (2020) for an introductory article in corpus linguistics;e and Larson-Hall and Herrington (2009), LaFlair et al. (2015), or Plonsky et al. (2015) for introductory treatments in the context of applied linguistics. Accordingly, you write a small script in R that does the following 1000 times: You take a random sample of words from the corpus with replacement and, then, you count the occurrences of *give* and *given* in them and store those frequencies in two vectors, one for each verb form. Next, you use those 1000 element vectors to compute the 2.5% and 97.5% quantiles for both *give* and *given* to obtain the confidence intervals of the bootstrapped frequencies. The analysis reveals that your findings are very similar to the results of R functions such as prop.test/binom.test and, thus, again support your hypothesis:

- 95% $\text{CI}_{\text{bootstrapped}}$ of *give*: [405, 484];
- 95% $\text{CI}_{\text{bootstrapped}}$ of *given*: [327, 401].

However, the problem still persists because the CIs computed as above assume a binomial distribution and/or a complete independence of data points (i.e., the computation assumes the so-called bag-of-words model, according to which corpora are “unstructured bags of words”); this implies that this approach does not consider the division of the corpus into, here, 500 parts/files (or 13 sub-registers or 5 registers), which indicates that this approach does not consider that the probability for any word to show up more than once in one text is higher than the probability of the word to show up more than once in a corpus as a whole (see the fitting sub-title of Church’s famous paper in 2000 “The chance of two Noriegas is closer to $^{p}/_{2}$ than p^2.”). Why does this matter? It matters because this systematically distorts the computation of the CIs and the difference between the kinds of CIs just mentioned – the first parametric ones from the functions prop.test/binom.test, the second ones from the bag-of-words-based bootstrapping), and, third, the more appropriate text/file-based bootstrapping ones that ‘respect’ the division of the corpus into parts) – can ‘make or break’ an analysis (depending on one’s hypotheses and significance thresholds, obviously).

Consider Fig. 1: The left panel represents the results one obtain from the inappropriate bootstrapping based on the corpus-as-unstructured-bag-of-words model. Red results pertain to *give*, blue results to *given*, the histograms show the distribution of the bootstrapped frequencies of *give* and *given,* and the lines/numbers at the top represent the CIs and their limits. Importantly, in the left panel, the red and blue confidence intervals do not overlap, indicating a significant difference between the frequencies of *give* and *given*.

The right panel represents the corresponding results from the appropriate bootstrapping over texts/files, and here, the CIs do overlap; in fact, the overlap of the blue results is so massive that ≈50% of the blue bootstrapped values for *given* are of a kind that are observed for the red bootstrapped values for *give* and vice versa, which makes it problematic to consider that the original hypothesis is supported.

This little case study demonstrates two things that lay the foundation for this largely programmatic paper (see Section 4 for the more applied case study). First, it demonstrates that the very widespread approach to report frequencies in corpora and then inferring something from that can be problematic. It is just not good practice to provide a summary statistic (like an overall frequency or mean) without a corresponding measure of variability or statistical dispersion. All studies basing their interpretation on, for instance, the difference of two frequencies as shown in the above example are prone to be inaccurate. I do not intend to argue that their conclusions *are* in fact wrong, but that, depending on the size of the difference and other characteristics of the data (e.g., their variability), we cannot know whether they are correct or not.

Second, it demonstrates that, even if that first point was conceded, one must recognize that not all measures of variability or statistical dispersion are equally good. Many studies to date have argued that the bag-of-words model is inadequate (Church, 2000; Evert, 2006; Lijffijt et al., 2011). Consequently, regular parametric approaches (e.g., based on binomial distribution) to CIs are problematic, which includes the "standard" prop.test or binom.test applications; similarly, bootstrapping with the word as a sampling unit is also problematic. More precisely and as noted above, the problem is not that **point estimates**, i.e., the me(di)ans, resulting from bootstrapping are problematic – in the above case study, all the results nicely converged around the observed frequencies of 444 and 365 for *give* and *given*, respectively. Rather, it is the **distribution** of the frequencies as reflected in the shape/width of the histogram that is quite different. The distribution of frequencies can have profound implications for the analyst's decision and interpretation because it is the shape/width of the histogram that would end up determining an analyst's decision of what is significant. However, bootstrapping approaches with linguistically more meaningful units of texts fare much better, and this is indeed the logic discussed first (for general corpus linguistics) by Gries (2006), who used bootstrapping on the basis of files, sub-registers, and registers, and then adopted in, for instance, Lijffijt et al. (2011, 2015), as well as popularized in some excellent overviews such as LaFlair et al. (2015), Plonsky et al. (2015), or Egbert and Plonsky (2020) (see also (Gries 2021a) for several examples of bootstrapping and permutation tests).

While bootstrapping is now more widely known in corpus linguistics than, say, 15-20 years ago, it seems that, except for Gries (2006), it is usually discussed as a complement to, and/or replacement of, more traditional parametric statistical methods such as linear models (e.g., *t*-tests or ANOVAs, see especially LaFlair et al. (2015) or Plonsky et al. (2015)), because the distributional assumptions of such traditional parametric methods are often violated by the Zipfian distributions so characteristic of corpus data.[2]

However, as may have already become clear from the introduction, this paper extends some of the arguments of Gries (2006) to make the point that most fundamental corpus statistics – not only frequencies, dispersions, and associations, but also others – benefit from having their uncertainty/variability quantified with bootstrapping approaches (or simulation-based approaches more generally). Put differently, simulation-based approaches such as bootstrapping do not just help with *t*-test type statistics, but they already help at an earlier stage, namely when we generate the kind of data for *t*-tests. Section 2 of this paper exemplifies the use of file-/part-based bootstrapping to exemplify the quantification of the uncertainty/variability that accompanies the combination of (i) observed frequencies and (ii) dispersion values (a normalized version of the Kullback-Leibler divergence (D_{KL}); see below for more information and Gries (2020) for an example). Section 3 does the same for association, quantifying the uncertainty/variability that accompanies the combination of (i) observed frequencies and (ii) again D_{KL}, a measure of association (see Baayen 2011 for the first mention of this possibility to the best of my knowledge and Gries and Durrant (2020) for discussion in a recent overview) that is less correlated with frequency of co-occurrence than other, more widely used association measures such as G^2 or *t*. Section 4 then discusses a practical application of the use of these methods based on COHA data analyzed for an amicus brief to the Supreme Court of the United States on the diachronic use of *gender* and *sex*. Section 5 concludes.

2. Frequencies and dispersions

2.1. Methods

To exemplify the computation of uncertainty/variability estimates/intervals through bootstrapping, imagine a scenario where you study a variety of verbs from an intermediate frequency range (e.g., between 100 and 10,000 pmw) to assess differences in the 'commonness' of their forms in a corpus such as the ICE-GB. I am putting *commonness* in single quotes here to indicate that I am using it as a technical term, one that is related to, yet distinct from, mere frequency of occurrence. A word *w* is 'common' in the most prototypical way if most or all of the following conditions hold:

[2] The notion of "Zipfian distribution" for corpus data refers to the fact that the frequencies of words in corpora in general or the frequencies of words in, say, lexically, grammatically, or constructionally defined slots in particular exhibit a distribution such that (i) a very small number of word types are highly frequent and (ii) a very large number of word types are extremely infrequent or even hapaxes. For example, of all the word forms between angular brackets in the ICE-GB, (i) the 30 most frequent word types (out of all 58,309) already account for 38.8% of all tokens in this 1 million words corpus and (ii) ≈58% of all word types each occur only a single time in the corpus. In other words, a Zipfian distribution is a power law function; see Manning & Schuetze (1999:20-29) for a more theoretical discussion of Zipfian distributions in corpus data and Ellis et al. (2016: Sections 2.2.3, 3.3.1, and 7.3.1, to name but a few) for more comprehensive exemplification.

(1) an average speaker of the language is extremely likely to know and use *w*;
(2) an average speaker of the language has known *w* for a long time (they learned *w* at an early age);
(3) an average native speaker of the language is extremely likely to encounter *w* regularly, which means
  a. *w* is sufficiently frequent;
  b. *w* is sufficiently dispersed in the language; and
(4) a non-native speaker of the language is likely to learn *w* early in their learning/acquisition process.

With this list, I am not implying that these are new criteria – they are not: Criteria 1 and 3a are operationalizable by frequency of occurrence in general corpora, criterion 3b is operationalizable by dispersion in general corpora; criterion 2 is essentially age-of-acquisition; and criterion 4 is operationalizable by frequency of occurrence and dispersion in learner corpora. I am merely introducing 'commonness' here as a notion that is intuitively straightforward to grasp and yet is defined here with a variety of generally well-understood measures and, thus, goes beyond what much linguistic work seems to use to represent 'commonness', namely just frequency of occurrence.[3]

Imagine further that, for this case study, we operationalize 'commonness' as a combination, but not conflation, of the two values of frequency and dispersion. That is, for each (case-insensitive) verb form of interest, we will compute its Zipf scale frequency for the ICE-GB, its dispersion in the ICE-GB (using D_{KL}), and uncertainty estimates based on bootstrapping over texts/files for both those values. The most straightforward and computationally efficient approach to explain and do all this is based on a **term-document matrix**, which contains in each cell the frequency with which the word form of that cell's row is attested in the text/file of that cell's column. If we define a word form heuristically as any case-insensitive string between angular brackets of the lines of the ICE-GB, the term-document matrix has 58,309 rows and 500 columns. As a corollary of the above-mentioned Zipfian distribution of word frequencies, this matrix is extremely sparse: 98.9% of these 58,309 × 500 cells are 0, as shown here for four words and only the first three files:

```
##            S1A-001 S1A-002 S1A-003
##   the          113      42      61
##   try            1       0       0
##   table          0       0       0
##   tolstoy        0       0       0
```

The row sums of this term-document matrix correspond to the word forms' observed frequencies, which we can convert to simple **Zipf scale frequencies** following Van Heuven et al. (2014, see esp. p. 1179f.); for simplicity's sake, I am using the formula that does not involve an adjustment for unseen words:

$$Zipf\ scale = 3 + log_{10} frequency_{pmw}$$

Conveniently, this formula already involves a log (which is added at a later stage in many applications of frequency values); this means that words that occur once in a corpus with exactly 1m words have a Zipf scale frequency of 3, whereas words that occur 1000 times in a corpus with exactly 1m words have a Zipf scale frequency of 6. This formula is applicable to the frequencies of the above four words in all 500 files:

```
##     the     try   table tolstoy
##   58614     335      74       1
##      the      try    table  tolstoy
## 7.738897 5.495941 4.840128 2.970896
```

[3] For example, Duyck et al. (2004) state that "[i]t is important to control for word frequency in psycholinguistic experiments because this variable has subtle effects, emerging not only between highly frequent and highly infrequent words, but even between frequent and slightly less frequent words". Similarly, Baayen et al. (2016) study the role of word frequencies in psycholinguistic work, and their following statement also implies that word frequencies are widely used in psycholinguistic work: "An assumption that lies behind the use of corpora in much psycholinguistic work is that a suitably representative corpus of, say, English can serve to represent (or control for) subjects prior lexical experience in accounting for various aspects of linguistic behavior" (p. 6); they then point to an additional important aspect to be considered when word frequencies are used, namely that "in using frequency counts for the study of specific aspects of lexical processing it is important to consider the communicative goals of the texts sampled by a given corpus and the specific demands imposed by a given task probing aspects of lexical processing."

The column sums of this term-document matrix correspond to the file/part sizes, which can be converted into relative file sizes (a numeric vector we might call file.sizes.rel in R) by dividing them by the overall corpus size, here shown for the first three files; in the ICE-GB with its 500 very similarly sized files, those are all fairly close to 2000:

```
## S1A-001 S1A-002 S1A-003
##    2192    2162    2280
```

```
##
##     S1A-001     S1A-002     S1A-003
## 0.002049918 0.002021863 0.002132214
```

With all this information, we can now compute the dispersion of each word form. The **Kullback-Leibler divergence** (D_{KL}) is a directional measure of how much a probability distribution $P_{1\text{-}n}$ diverges from another probability distribution $Q_{1\text{-}n}$ (i.e., it is not a symmetric distance metric). In applications of the D_{KL},

- the probability distribution $P_{1\text{-}n}$ is usually the one that reflects a posterior distribution and/or an observed distribution, i.e., data.
- the probability distribution $Q_{1\text{-}n}$ is usually the one that reflects a prior distribution and/or a theoretical or an expected distribution.

In our application here, Q corresponds to the file sizes, the numeric vector we called file.sizes.rel above, which states for each file the proportion it makes up of the whole corpus. A word can be considered as completely evenly distributed across the 500 corpus parts if its relative frequency in each corpus part corresponds to the proportion that each corpus part makes up of the whole corpus. P, on the other hand, contains the proportions of occurrences of a word in a corpus part. We have seen above that *table* occurs 74 times in the corpus, and 7 of these instances are in the last file (W2F-020), meaning the value for *table* in W2F-020 is $^{7}/_{74}$=0.0945946. The following formula shows how D_{KL} is computed:

$$D_{KL}(P||Q) = \sum_{i=1}^{n} P_i \times log_2 \frac{P_i}{Q_i}$$

However, because we already generated a term-document matrix above, computing all 58K D_{KL}-values takes only one function call (e.g., in R using apply to a function that computes D_{KL}s to all rows of a term-document matrix). As a divergence, D_{KL} falls into the interval [0, +∞] and essentially quantifies how much information in bits (because of the log to the base of 2) we lose if we try to approximate P with Q. However, to make it more straightforward to compare D_{KL} values from differently long probability distributions, we can normalize D_{KL} to fall into a [0, 1] interval as follows, which also has the side effect that this measure is nicely correlated with, for instance, Gries's DP_{norm} ($r = 0.928$):[4]

$$D_{KLnorm} = 1 - e^{-KLD}$$

The D_{KLnorm} values for the four example words from above are given as follows:

```
##       the       try     table   tolstoy
## 0.1011116 0.8020354 0.9772823 0.9998619
```

If we plot dispersion against frequency, we obtain the distribution that is characteristic of most general corpora and most dispersion measures, as shown in Fig. 2.

For the exemplification of the bootstrapping approach, I will compute bootstrapped frequencies and dispersions for all 179 verb types in a moderate frequency range (frequencies between 100 and 10,000 in the ICE-GB), but I will focus the more qualitative discussion on six verb lemmas from that range: COME, GIVE, KNOW, LOOK, MAKE, and TAKE. I define the number of bootstrapping iterations as 1000 and perform the following steps for each of these iterations. Thus, we

- Sample 500 files names with replacement.
- Retrieve the words and the file names for these 500 files (with repetitions as needed) and create a term-document matrix.
- Compute
  - The frequencies of the 179 words in the 500 bootstrapped files.
  - The dispersions of the 179 words in the 500 bootstrapped files.
- Store the results for the frequencies and the dispersions in two separate "collector" matrices, which
  - Have 179 rows: one for each verb type included in this little application.
  - Have 1000 columns: one for each iteration.

[4] A reviewer suggested to use 2 and not e as a base in the normalization; this, like other transformations (e.g., min-max), is perfectly possible and generates values that are mathematically/deterministically related to the ones used here and, while they will change the exact numerical results, they will do so perfectly predictably without changing the overall argument of the paper (that we should be providing (file–/text-based bootstrapped) uncertainty estimates for our corpus statistics). Moreover, I am using D_{KL} rather than another dispersion measure because we can then use the same measure both in this dispersion case study and in the association case study – this would not be possible otherwise and, thus, introduce more complexity with no additional advantages.

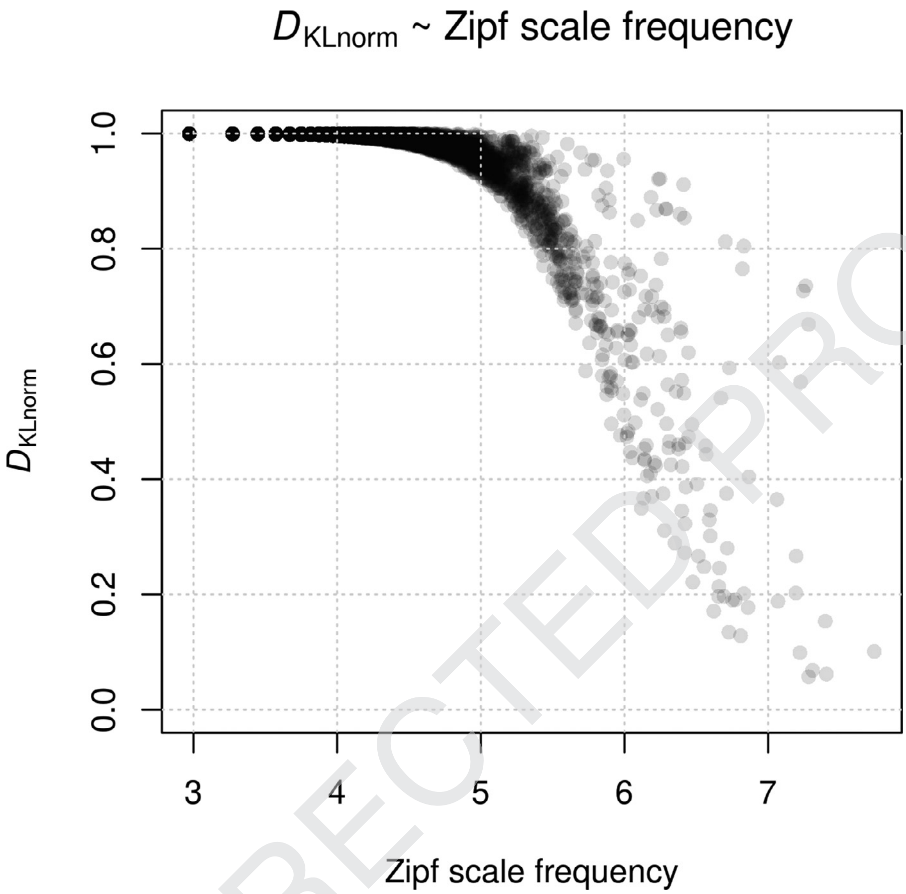

Fig. 2. The correlation of frequency and dispersion (as measured by D_{KL}).

- Contain in each cell the frequency or dispersion of the relevant word in the relevant iteration.[5]

A small part of the collector matrix with the frequencies is shown below. In the first bootstrapped version of the corpus, *accept* occurred 119 times, but in the fourth one, it occurred only 92 times, indicating that there is some variation:

```
##          ITERATIONS
## VERBS      1   2   3   4   5   6
##   accept 119 108  97  92 102 107
##   agree  160 171 155 172 147 146
##   agreed 135 116 114 107 132 107
```

2.2. Results

Given the variation shown above (e.g., for *accept*), let us first determine how much the bootstrapped results vary and how much that variation relates to the frequencies and dispersions of the word forms. Because we already have the frequencies and dispersions,

[5] An R script to convert the ICE-GB heuristically into an R data frame, compute frequencies and dispersion for all word types, and compute CIs in the three ways discussed here (parametric, bootstrapping in the bag-of-words model, and the proper file-/text-based approach) is available at <https://www.stgries.info/research/2022_STG_UncertEstimates4CorpStats_RMAL.zip>.

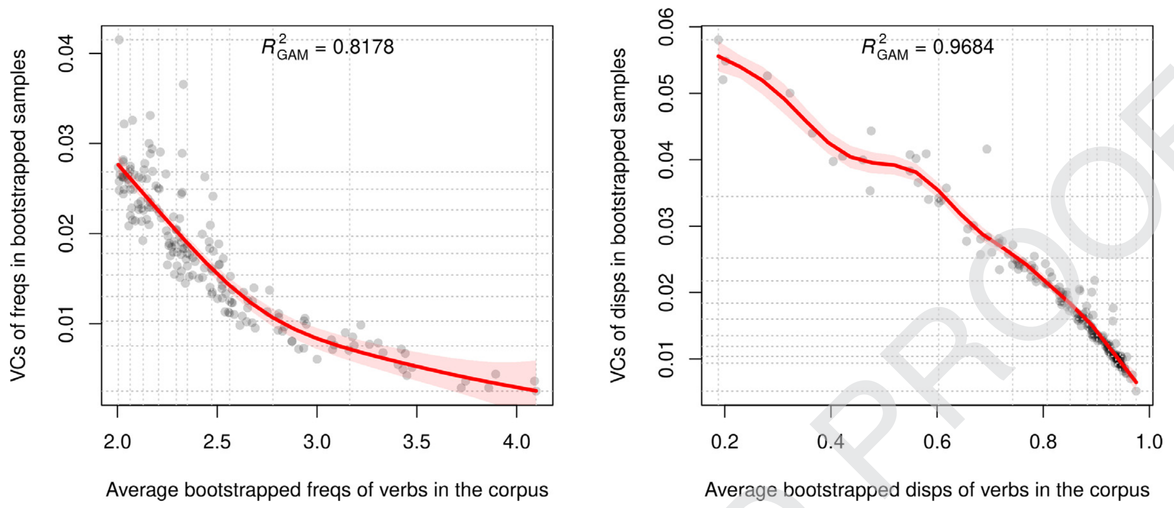

Fig. 3. Correlations of uncertainty with frequency and dispersion.

we only need to quantify their variation using the variation coefficient. The variation coefficient of a vector of numbers is the standard deviation of that vector divided by the mean of that vector, as shown here on the basis of a simple vector with numbers from 1 to 5:

```
##    standard deviation                   mean       variation coefficient
##             1.5811388              3.0000000                   0.5270463
```

The variation coefficient has the advantage over the standard deviation in that it quantifies variability while taking the mean of the distribution into consideration. In other words, it quantifies the variability of the numerical values relative to their mean (see Sheskin 2011:15f.). In this case, where the frequencies of the 179 verbs differ considerably because they were sampled from the frequency range of 100 to 10,000 (i.e., 100 times as many), comparing the standard deviations without controlling for those differences is sub-optimal. Thus,

- I computed for each verb
  - The mean and the standard deviation of its 1000 frequencies in the bootstrapped samples.
  - The mean and the standard deviation of its 1000 dispersions in the bootstrapped samples.
- I computed each verb's variation coefficient for the frequencies and dispersions.
- I represent those graphically in two scatterplots in Fig. 3 (with grid lines for indicating deciles) and summarize them with a regression line (plus confidence band) from a generalized additive model.

This demonstrates that quantifying the uncertainty of our corpus statistics becomes more important as the words in question become less frequent and less evenly distributed. For the most frequent and/or evenly distributed word forms, there is not much variability, but the variability increases as the frequencies/dispersion of the word forms decrease.

Let us do some plots of six lemmas of interest. For interpretability's sake, this will be done in three plots with two lemmas each. Each plot shows all forms in grey and then highlights the results for the two lemmas of interest in blue and red. In all three plots, the x-axis is form frequency logged to the base of 10, the y-axis is D_{KLnorm}, and the shaded area around each form of interest represents 90% data ellipses (the R function car::dataEllipse, version 3.0-11, see Fox & Weisberg 2019) that summarize the distribution of the 1000 sampled values for each number with "estimated probability contours, containing expected fractions of the data" (Fox & Weisberg 2019: Section 5.2.3). First, LOOK and MAKE; second, COME and GIVE; third, TAKE and KNOW. The results show

- For MAKE, a "clustering" of {*makes*, *making*} vs. {*made*, *make*} for both frequency and dispersion.
- For LOOK, a "clustering" of {*looks*, *looked*} vs. *looking* vs. *look* for both frequency and dispersion (the latter two nearly overlap for dispersion, but not quite).
- For GIVE, a "clustering" of {*gives*, *giving*, *gave*} vs. {*given* *give*} for both frequency and dispersion (but only because of the wide ellipse for *giving*).
- For COME, a "clustering" of {*comes*, *coming*} vs. *come* for both frequency and dispersion.
- For KNOW,
  - in terms of frequency, there is only one cluster (but only because of the wide ellipse of *knew*).
  - a dispersion ranking of *known* < {*knew*, *knows*} (with a very small distance between *known* and *knew*).
- For TAKE,
  - a frequency clustering of *takes* ≤ {*taking*, *took*} < *taken* < *take*; but
  - a dispersion clustering of *take* < *taken* < {*took*, *taking*} < *takes*.

Note that the clusters for these verbs are very similar across frequency and dispersion because the dispersion measure used here – D_{KLnorm} – is correlated with frequency in general (even if less so than other dispersion measures (see Gries, to appear, for a correlation between frequencies and different dispersion measures in a few corpora; in that study, of the more widely used dispersion measures, D_{KL} is least correlated with frequency). If these calculations were performed with a dispersion measure specifically designed to be less correlated with frequency, this could change (see again Gries, to appear, on how such a dispersion measure could be developed).

2.3. *Interim discussion*

The results are clear. For each of the verb forms investigated, we find that the exact frequency and dispersion value computed from the corpus exhibit a sizable amount of variability on both dimensions of frequency and dispersion. A study whose implications rested on how the forms of a lemma were ranked in terms of their commonness (or even just frequency and dispersion separately) would have to concede that, even if the results for the whole corpus supported every hypothesis, they show such a high amount of variability that many of the form-by-form rankings must rather be seen as consisting of forms that are not significantly different for both frequency and dispersion; this is particularly obvious for the forms of KNOW included here.

The proposed approach has several appealing characteristics. First, we can evaluate frequency and dispersion differences for significance in a way that avoids the problem of either of the two ways in which this is usually done in corpus-linguistic practice. On the one hand, many studies have compared the frequencies of elements (morphemes, words, *n*-grams/lexical bundles, constructions, and other elements) only with reference to the elements' absolute observed frequencies, but we can see that this is problematic given how volatile such absolute frequencies/dispersions can be. On the other hand, several studies have used a chi-squared/log-likelihood test to support claims about how 'robust' the difference between some frequencies might be (see learner corpus studies such as Aijmer (2005), Altenberg (2005), Connor et al. (2005), Laufer and Waldman (2011), Gilquin and Granger (2011), Neff van Aertselaer and Bunce (2012), and keyword studies such as Hofland and Johansson (1982), Leech and Fallon (1992), Scott and Tribble (1996), Culpepper and Demmen 2015). However, strictly speaking, neither of these tests is permitted because both are based on the wrong bag-of-words model and violate the assumption that all data points are independent of each other. The present approach avoids both problems, and we can clearly see the extent of uncertainty/variability for both the frequencies and dispersions, which allows us to make more substantiated statements and comparisons regarding commonness, frequency, and dispersion.

Second, this approach allows us to simultaneously and separately consider the uncertainty of frequency and dispersion. Rather than conflating frequency and dispersion into an adjusted frequency index, which is sometimes done in lexicography (see, e.g., Gardner & Davies 2014) and which, because of the information loss it causes, often returns results that can be highly misleading, the present approach of (i) keeping both dimensions of information and (ii) quantifying their uncertainty gives us a clear picture of words' distribution in a corpus/corpora. In fact, the present bootstrapping approach, while first developed in Gries (2006), is conceptually similar to Egbert and Plonsky's (2020:600) discussion of (the problems of) vocabulary lists and how they could be improved. Specifically, they propose

> An alternative approach would be for researchers to use bootstrapping to collect many resamples of texts from the corpus, by storing a word list each time. The simplest way of creating a word list using these resampled word lists is to include any word that occurred in at least N% of the resampled lists. This approach combines the measurement of frequency and dispersion into a single step. It is a much more robust method that is less dependent on the design and contents of the original corpus sample.

However, I submit that the present approach goes beyond this suggestion in two ways by

- not just using *range* as a measure of dispersion like their proposal implies but a measure that also takes frequencies of words in corpus parts and the sizes of corpus parts into consideration (while avoiding Juilland's *D*, which some newer studies have found to be problematic (Biber et al., 2016, Burch et al., 2017).
- not just using the bootstrapping result as a way to identify a cutoff point (although that is also a possible application) but also representing the degree of uncertainty/variability to inform cutoff points as well as rankings or other research decisions.

Finally, all this is achieved on the basis of a linguistically meaningful sampling unit, namely the file, which in this corpus represents the text or linguistic interaction. This is important because not respecting these linguistically meaningful divisions in our corpora, such as by dividing corpora artificially into *n* equally large parts even if that cuts across texts and/or subregisters, has been shown to lead to suboptimal results (Burch et al., 2017). Here, however, the bootstrapping was done on files (in Gries (2006), it was done on files, sub-registers, or registers), which preserves the integrity of the sampling process by not sampling across text boundaries. Let us now apply the same logic to the study of association measures.

3. Association

3.1. *Methods*

To exemplify the text/file-based bootstrapping approach in the domain of research on co-occurrence/association, I will use a collostructional case study asking how much words are attracted to a slot in a construction. As an example, I will use the ditransitive construction simply because it is well known to make for a good test/validation case. As mentioned above and slightly tweaking Baayen's (2011) suggestion, I will use D_{KLnorm} as an association measure, which has two advantages:

- It is less strongly correlated with the raw frequencies of the elements involved than some other widely-used association measures, such as G^2 or t, which really mostly reflect co-occurrence frequency rather than association (see Gries, to appear a, for empirical evidence to that effect).
- It is a directional measure, allowing us to focus on one direction of association rather than consider only mutual association. The direction of association I will use is from the verb to the construction (v2c).

After computing D_{KLnorm} for dispersion, where we compared the frequencies of words in files against the file sizes, the question arises of how the D_{KLnorm}(v2c) is computed for association. Following Baayen (2011), it is based on the same 2×2 co-occurrence table as most other association measures and, for the current example, it involves computing the (normalized) divergence of the percentage distribution in the first row of this 2×2 co-occurrence table from the percentage distribution of the column totals. In this case, the observed distribution of *give* across ditransitives and other constructions (0.5337838 vs. 0.4662162) is very different from the proportion of the ditransitive in general (0.0802808 vs. 0.9197192); hence, we obtain a rather high value for how much the construction is attracted by the verb form *give*:

```
##       DITR
## GIVE    yes    no   Sum
##   yes  237   207   444
##   no  1604 20884 22488
##   Sum 1841 21091 22932
## KLD norm. from give to ditrans
##                      0.6328251
```

As above, let us first check the non-bootstrapped values for all verb forms that result from applying this to every single verb form ever attested in the ditransitive in the ICE-GB, as shown in Fig. 7.

The results are encouraging: Forms of very prototypically ditransitive verbs score high on both dimensions. The rightmost forms (those most often occurring in the ditransitive) and the topmost forms (those most attracted to the ditransitive) are forms of the lemmas GIVE, TELL, SEND, ASK, OFFER, etc. We also see that D_{KLnorm} is, as desired, much less dependent on frequency of co-occurrence: With the exception of forms occurring more than 2^6=64 times in the ditransitive, the attraction of the forms to the construction is not obviously predictable from the co-occurrence frequency, which means that each dimension contributes something largely unique to the analysis.

To exemplify bootstrapping, we will proceed in a way that is analogous to the one above. We perform the following 1000 times:

- Sample 500 file names with replacement.
- Cross-tabulate for all verb form tokens when they are used in a ditransitive and when they are not.
- Compute D_{KLnorm} (v2c) for each verb form and store the results in a collector structure.

3.2. Results

We plot the verb forms again by using data ellipses as in Figure 8.

We observe that the verb forms that score the top slots, as in Stefanowitsch & Gries (2003:299), are of the lemmas GIVE and TELL. Interestingly, the forms of TELL seem to score slightly higher on association than the forms of GIVE, but we can also see that there is some overlap of the ellipses. Because most collostructional analyzes (and many collocational studies) use lemmas, I changed every verb form into its lemma and redid all calculations. The results are shown in Fig. 9, with several verbs highlighted:

It is clear that GIVE and TELL have the most prominent positions and do not differ from each other either in terms of frequency or association. GIVE could be seen to score the top spot if one converts all logged frequencies and association scores to the 0-1 range (applied the min-max transformation) and compares the Euclidean distances of all verbs to the origin. Remarkably, both verbs make to the top of the list by outperforming all others in terms of frequency, but not association. We can also see that, in general, verbs differ more significantly in terms of frequency than in terms of association: the data ellipses are usually much higher than they are wide. As a result, many of the “usual suspects” – SEND, SHOW, OFFER, for instance –do not differ reliably from each other, especially not on the dimension of association. Finally, the two huge ellipses on the left are for ASSURE and REASSURE, showing that these two morphologically related verbs are extremely similar in their attraction to the ditransitive and could for most analytical purposes be conflated.

3.3. Interim discussion

While we only considered a few verbs here, it is clear that, just like frequencies and/or dispersions, rankings of association measures as they are often used in corpus linguistics are also a little less straightforward than is traditionally assumed. This is for two reasons:

1. Many traditional association measures – G^2, t, z, p_{FYE}, MI^2 – conflate the main two dimensions of information they are computed from – co-occurrence frequency and association – into one dimension, and while that can yield instructive rankings, it is not guaranteed to do so especially because the degree to which the rankings of association measures is in fact more influenced by frequency, not association, can be extremely high (see Gries, to appear a, for an entire paper on this problem). As mentioned

above, the approach discussed here allows us to include both dimensions separately and in a visually easily comprehensible manner.

2. More importantly, any such rankings do not take into consideration the volatility of the measures or, put differently, the (apparently quite high) degree of dependence of the (rank of the) measure on the exact composition of the corpus where, as in the earlier section on frequency and dispersion, the present approach bases its sampling not on the problematic bag-of-words approach but a sampling based on the linguistically more meaningful unit of a text.

Because of this, any inferences based on a ranking of association measures and any interpretation of the different values that verbs score should be more tentative/careful. For instance, as analysts we should probably make less of "a big deal" out of the fact that GIVE ranks a bit higher than TELL on the combined association to the ditransitive (however much other semantic considerations might welcome that result) since we have now seen that

1. There is some spread between the different forms of the two lemmas (reminiscent of the forms vs. lemmas discussion of Rice and Newman (2005) and Gries (2011)).
2. The association measures used most often do not just convey association, but they also convey frequency (and often much more so).
3. The ranking results are quite dependent on the corpus being exactly as it is such that even minor changes in the corpus composition can affect rankings considerably.

It is worth pointing out that it is possible to add a third dimension – dispersion – and its uncertainty/data ellipse to the mix to generate a three-dimensional version of Fig. 9 with uncertainty spheres. However to keep complexity manageable, I am not doing this here. Let us now consider a diachronic case study from the domain of legal/forensic linguistics.

4. The diachronic development of *gender* and *sex* in COHA

4.1. Introduction and the "traditional approach"

I conclude this largely programmatic paper with one more concrete example of the above logic. In a recent amicus brief filed with the Supreme Court of the United States for three cases involving claims of discrimination based on sexual orientation and gender identity, Eskridge et al. submitted an analysis of corpus data on how the word *gender* was used in the 1960s, which was then also compared to how the word *sex* was used at that time. Amici suspected that *gender* was essentially a non-word in the 1960s when Title VII was enacted and became more common only later. If this hypothesis was correct and if speakers, therefore, used *sex* in the 1960s while speakers later and today would use *gender*, then one could argue that the original formulation of Title VII ("because of […] sex") can be understood as protecting gay and transgender people against discrimination.

One standard way of showing that (i) *gender* was essentially a non-word in the 1960s and that (ii) it became more common since then would be to determine the frequency of *gender* in a representative corpus of 1960s American English and then chart its frequency development since then to the present. In the above-mentioned amicus brief, the authors used the COHA data from the five decades involving the 1960s to the 2000s components of the corpus. I will use the same data, but I will not conduct a concordance line-by-concordance line an alysis and disambiguation of cases; this is only for expository reasons, and the main methodological argument is not affected by that. I wrote R scripts that would retrieve from the tabular version of the relevant decades of COHA all instances of the lemma GENDER (combining *gender* and *genders*) and the lemma SEX (combining *sex* and *sexes*) together with in which of the sampled texts they occurred how often. Figure 10 shows the simple Zipf scale frequencies (Van Heuven et al., 2014) for each lemma in each decade:

The plots are suggestive and compatible with the above hypothesis: Both words increase in frequency, but *gender* is rarer and increases more over the five decades under consideration. However, there are two problems. First, counter to the spirit of this article, these frequency data do not come with any indication of their variability. We do not know how much different these frequency estimates are likely to be if the corpus slices looked slightly different. Second and connected to that, these data are likely insufficient to represent the two words'/lemmas' commonness because they pretend that frequency is the only/best measure for commonness and disregard the words' dispersions in each decade. Thus, this is especially problematic for *gender* in the 1960s, because if one does a line-by-line analysis of all examples of *gender* in the 1960s, Eskridge et al. (2021:1554, note 240) find that

> [a]ll examples […] in the relevant sense […] but two were from a single one of more than 10,100 corpus files. In other words, had *Journal from Ellipsia* [an avant-garde science fiction novel about genderless aliens written by a feminist author …] not been sampled, there might not have been a single obvious and countable example in a twenty-four-million-word corpus covering ten years of American English."

Given areas where a traditional analysis can be improved, the next section will discuss the application of the bootstrapping logic to these data.

4.2. The bootstrapping approach

To address both issues simultaneously, I used the following procedure, which was applied separately to each of the five decades. I took 3,000 random samples of the size of the number of corpus parts with replacement, and for each of these samples, I computed

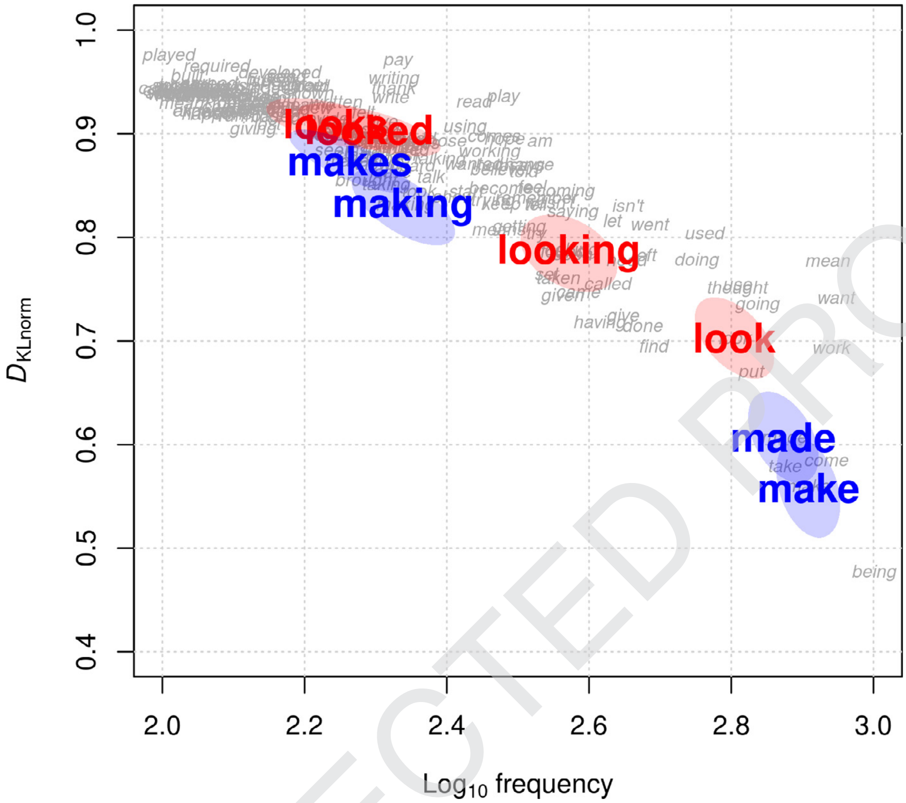

Fig. 4. The uncertainty of frequency/dispersion for MAKE and LOOK.

the Zipf scale frequency of the lemmas *gender* and *sex* as well as the range (the proportion of corpus parts that featured at least one instance of the lemma in question). The result was a data structure with 3,000 bootstrapped Zipf scale frequencies for each lemma and each decade and 3,000 bootstrapped ranges for each lemma and each decade.

Then, these were plotted in the same way as in Figs. 4–6 above:

- The *x*-axis represents the Zipf scale frequencies.
- The *y*-axis represents range logged to the base of 2.
- The numbers 6, 7, 8, 9, and 0 represent the mean frequencies and ranges for the decades 1960, 1970, 1980, 1990, 2000, where *gender* is represented in red and *sex* in *blue*;[6]
- The shaded areas around the points/numbers represent 90% and 95% data ellipses.

We can see that

- *Gender* is increasing in what seem to be about three stages with an altogether notable increase in both *gender*'s frequency and dispersion:
  - In terms of frequency (*x*-axis), there are two to three stages: {1960, 1970 (and maybe 1980)}, maybe {1980}, and {1990, 2000}, as indicated by the increases, and by the facts that the 1960s frequency ellipsis is so wide that it overlaps with both the 1970s and 1980s ellipses (which do not overlap with each other) and that the ellipses for 1990 and 2000 overlap as much as they do.
  - In terms of dispersion (*y*-axis), there are four relatively clear stages: {1960}, {1970}, {1980}, and {1990, 2000}.
- *Sex* is increasing in what seem to be two to three stages with an increase that, compared to the development of *gender,* seems relatively minor:
  - In terms of frequency, there seems to be a relatively continuous increase: the 1960s are lowest, followed by {1970, 1980}, followed by {1990, 2000}, but the relative continuity of the increase is because the 1970 ellipsis "holds together" the earlier and the later frequencies so that there are no more abrupt/distinct developmental stage/break.
  - In terms of dispersion, there are three clear stages: {1960}, then {1970, 1980}, then {1990, 2000}.

[6] I am not representing the decade of the 1960s with "60" or even "1960" (as suggested by a reviewer) because this would lead to massive overplotting and render the plot much less easily interpretable.

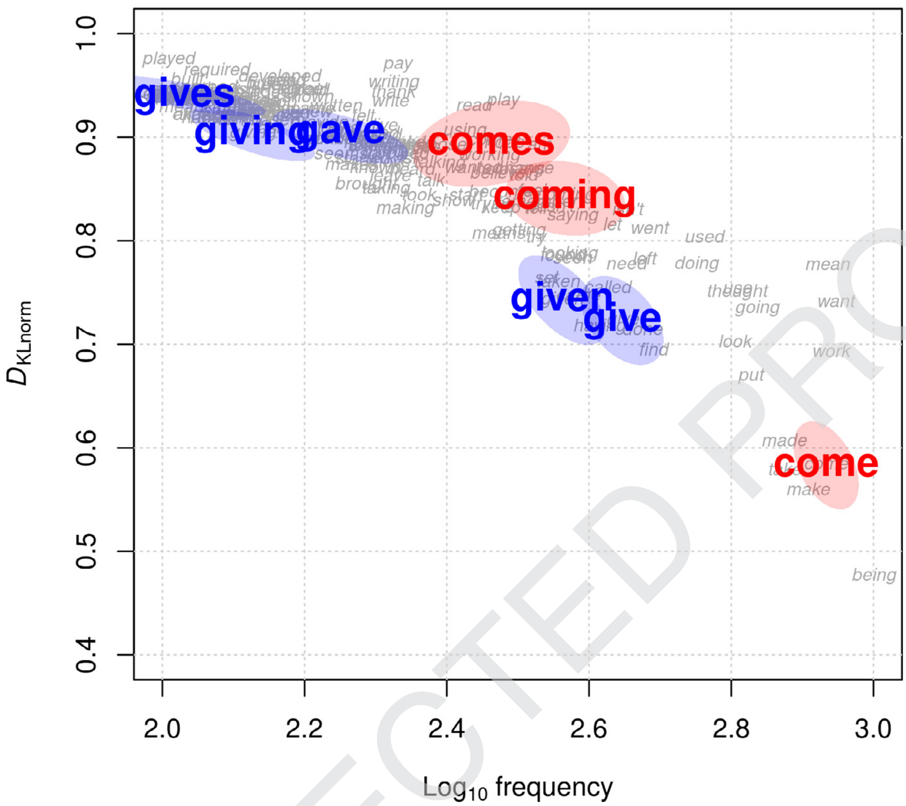

Fig. 5. The uncertainty of frequency/dispersion for COME and GIVE.

These data are certainly compatible with the above hypothesis: *gender* is so rare and so clumpily distributed in the 1960s that, for most practical purposes, it may well be considered a non-word. However, frequencies are difficult to interpret intuitively and dispersions are even more so. Thus, to understand what such values “mean,” or correspond to, some expository strategy is required. Eskridge et al. (2021:1554) try to make the extremity of *gender*’s clumpiness comprehensible by showing that *gender* is as clumpily dispersed in the 1960s corpus data as are text-processing errors in the corpus such as *brilliantp250that* (which should be *brilliant that*, omitting the extraneous page number). A probably better strategy, though, because it would also take frequency into consideration and would be more in line with this data ellipse approach of the current study, is to compute which words are closest to *gender* (and *sex*) in a given decade in the above plot (and therefore in the data ellipse). To that end, for both *gender* and *sex*, I computed their Euclidean distances from each of the ≈315K word types in the 1960s decade of COHA.

The following are the 24 words closest to *gender* (listed in lower caps because the computations were performed in that way): *whelan*, *kermit*, *0.5*, *interception*, *elgin*, *gibbons*, *socializing*, *but –*, *landau*, *stubbs*, *brenda*, *buzzell*, *innovator*, *g.i.s*, *homeric*, *organisations*, *lumpur*, *worthington*, *niles*, *entrepreneurial*, *- is*, *nutritive*, *evansville*, and *4,000,000*. These words are not nearly as “exotic” as the words one obtains if one only considers dispersion (as in Eskridge et al., 2021). They are also “intuitively rare”. Many of them are proper names or numbers, some involve punctuation, and one involves a British spelling that is much rarer in an American corpus than its American-spelling counterpart (the American spelling *organizations* is ≈31 times more frequent than *organisations*). Another way of evaluating the words close to *gender* would be by just heuristically assessing their age of acquisition, and except for *kermit* and the first name *brenda*, none of these words strike me as particularly early in acquisition by children during L1acquisition or by learners during L2 acquisition/learning. However, the picture we get from the words closest to *sex* are quite different: *11*, *credit*, *term*, *indicated*, *standard*, *okay*, *lawyer*, *families*, *begins*, *designed*, *horses*, *jesus*, *wine*, *forms*, *mountain*, *movie*, *saturday*, *coast*, *discussion*, *limited*, *magazine*, *birds*, *prevent*, *puts*. Many words on this list are words that strike me as quite common/basic and, arguably, as words that would be acquired/learned early in an American context (especially *okay*, *families*, *begins*, *horses*, *jesus*, *mountain*, *movie*, *saturday*, *discussion*, *limited*, *birds*, and *puts*).

The impressionistic evaluation from the above argument is supported by a more rigorous evaluation. I looked up the above 24 words closest to *gender* on the one hand and *sex* on the other hand and compared the distributions of their ages of acquisition ratings

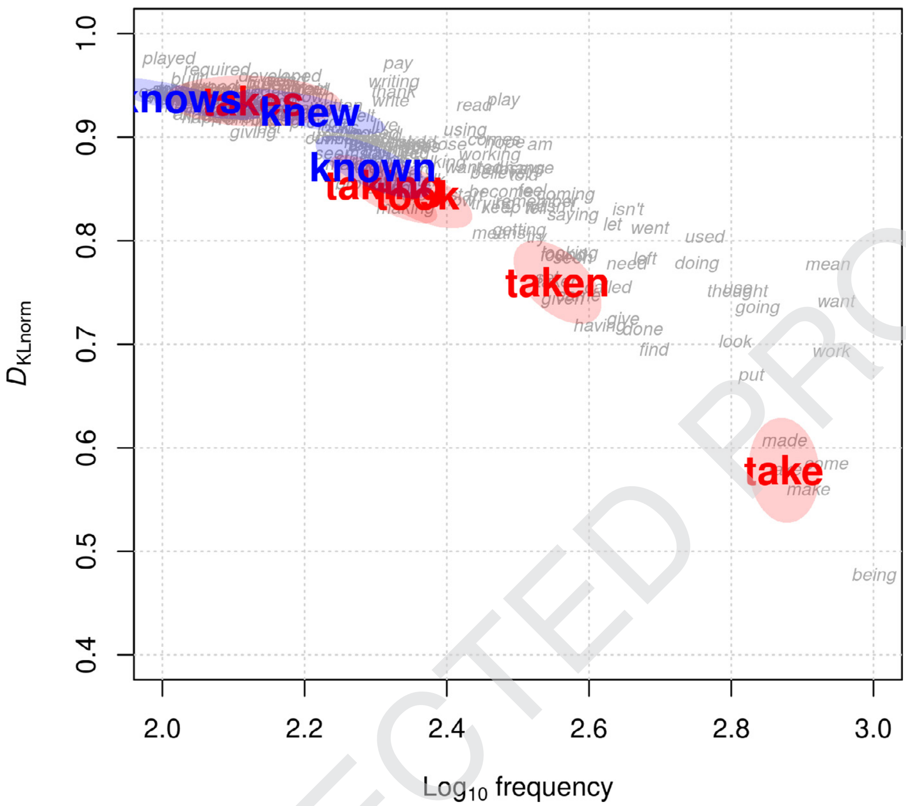

Fig. 6. The uncertainty of frequency/dispersion for KNOW and TAKE.

(based on Kuperman et al., 2012).[7] A two-tailed Kolmogorov-Smirnov test yielded a significant result (D=0.875, p<0.001) such that the words distributionally similar to *gender* have much higher AoA ratings than the words distributionally similar to *sex*.

Crucially, if one applies the same logic to the last decade covered here, the 2000s, the words closest to *gender* reflect how much more common this word has become over time: *assets*, *networks*, *faculty*, *genetic*, *clinic*, *wolf*, *designer*, *drivers*, *turkey*, *sandy*, *organic*, *allen*, *seeds*, *terrorists*, *executives*, *terrorism*, *chronicle*, *sheriff*, *mississippi*, *m.*, *korea*, *aids*, *massachusetts*, *alan*. It is true that not all words are words one would think every child or learner acquires or learns early (e.g., *faculty* or *chronicle*), but the number of words one might assume to be known to younger kids and earlier learners is still considerable (this seems to be true even for the names): *clinic*, *wolf*, *designer*, *drivers*, *turkey*, *sandy*, *sheriff*, *mississippi*, *korea*, *massachusetts*, *alan*. The AoA ratings for the words similar to *gender* in the 2000s fall right between, and are significantly different from both the AoA ratings for the words similar to *gender* in the 1960s and AoA ratings for the words similar to *sex* in the 1960s, as is also represented in the ecdf plot in Fig. 12.

From the frequency-cum-dispersion changes over time, their uncertainty ellipses, and the "changes in vocabulary similar to *gender/sex*" in the relevant decades, I would conclude that *gender* increased considerably in both frequency and dispersion from the 1960s onwards, as amici suspected in their brief and demonstrated there on the basis of a procedure less comprehensive than the one discussed here.[8]

5. Concluding remarks

I anticipate that the present paper succeeds in showing that the bootstrapping approach has attractive features. To reiterate, rather than just running bootstrapping as a replacement for parametric statistics, the examples discussed here show how text-/file-based bootstrapping

[7] The lookup was done in three stages: I first looked up the exact words listed above in the AoA ratings and used the AoA rating value if that word was included in the ratings. If the word was not included in the ratings, I looked up the lemma form and used that lemma's AoA rating instead (e.g., the singulars *horse* and *family* for *horses* and *families* or the infinitives *begin* and *design* for the forms *begins* and *designed*). AoA ratings for the words for which no AoA ratings were available were set to the highest AoA rating available for the noun in question (*gender* vs. *sex*) in the time period in question. I am grateful to Reviewer 1 for suggesting to use these ratings.

[8] This is not everything that can be said about how *gender* related to *sex* especially in the 1960s; see Eskridge et al. (2021) for much more detailed discussion (including concrete examples of 1960s uses of the *sex* that today would involve the word *gender* with expression such as *sex roles* and *psychosexual* or the relation to *transsexual* and *transgender*).

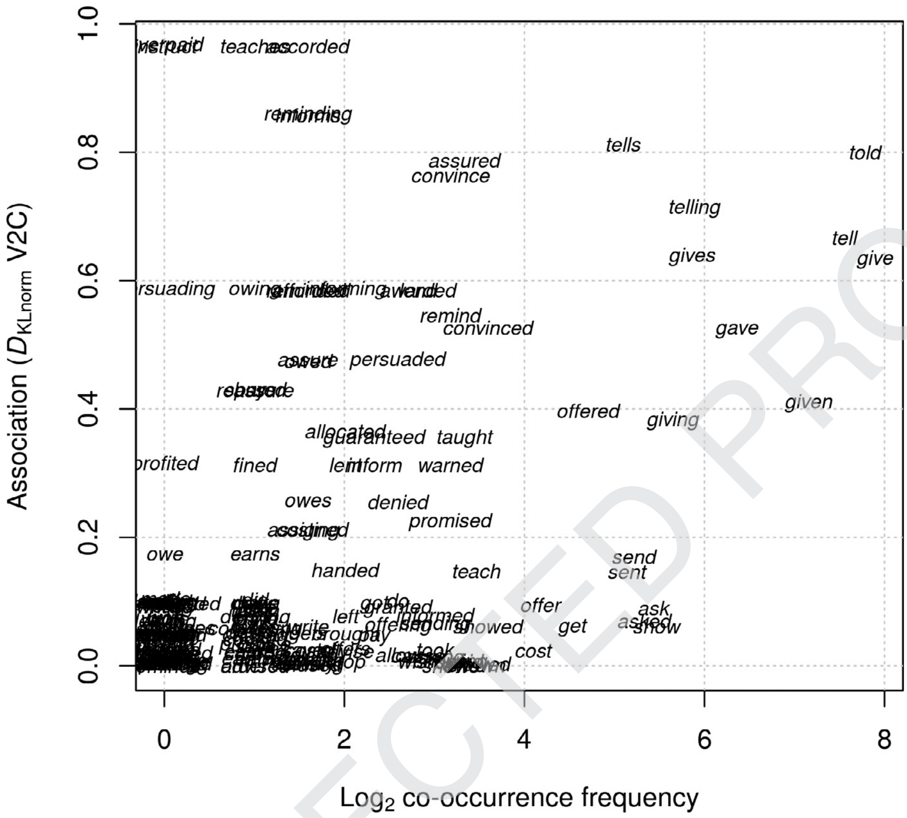

Fig. 7. The correlation between co-occurrence frequency and association.

Table 1
Zipf scale frequencies of *gender* and *sex* in COHA (1960–2000s).

Word	1960s	1970s	1980s	1990s	2000s
Gender	3.0432	3.1121	3.6617	4.23	4.3257
Sex	4.7207	4.8568	4.7796	4.9391	4.9742

- Can be used to generate the same point estimates as other approaches but statistically more appropriate and, thus, more reliable dispersion/distribution estimates for CI-like inference and interpretation.
- Can be extended to consider more than one metric at a time, as when we consider frequency-cum-dispersion or frequency-cum-association results.
- Can be represented in a visually straightforwardly interpretable way.
- Can allow all these to be achieved in ways that can be extended directly to other corpus-linguistic methods, such as keywords.

One reviewer questioned whether one needed bootstrapping for this and whether one can simply compute means and uncertainty estimates (such as standard errors and standard deviations) from the actual data. But we have seen in case study one in Section 1 above that the parametric CIs are anticonservatively narrow (given how they are based on the incorrect bag-of-words model and essentially rely on the binomial distribution or approximations). As Egbert and Plonsky (2020:593) state, “Bootstrapping […] is often recommended for small samples and samples with unknown or non-normal distributions” (see Egbert & LaFlair (2018) for more discussion in the context of applied linguistics and DiCiccio and Efron (1996) for a statistical discussion highlighting the advantages of bootstrapping over traditional CIs). This is because means and standard deviations only work best when they are applied to fairly normally distributed data, but frequencies of words in corpora, corpus parts, or lexically/grammatically defined slots are usually Zipfian-, not normally, distributed, as discussed in Section 1. To provide a brief example, Fig. 13 shows the distributions of files sizes (in words) of the BNC (left panel) and the frequencies (in files) of the randomly chosen word *furniture* (right panel). Clearly, quantifying the uncertainty with a mean and standard deviation is not a good idea for the data represented in these histograms (Figure 13)).

There are other advantages as well. For example, the visual representation suggested above not only gives us more information than a mere plotting of a single point estimate for each (year, frequency, dispersion) tuple would offer, but it also provides a nice

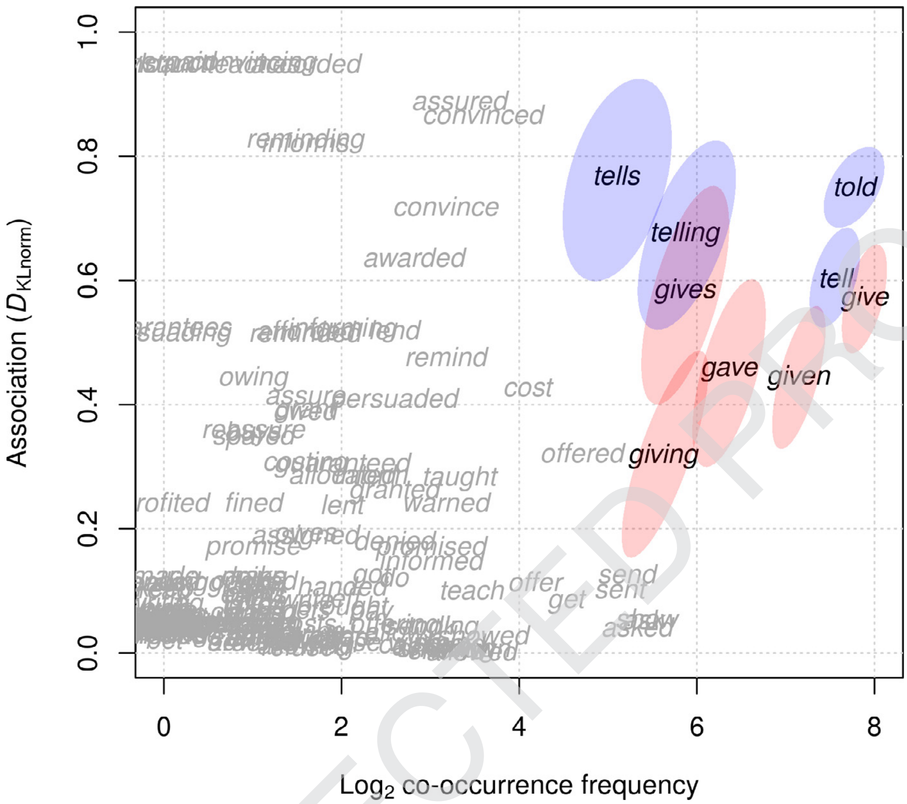

Fig. 8. The uncertainty of frequency/association for GIVE and TELL.

complement to work on identifying historical or developmental stages in corpus data. For example, Gries and Hilpert (2008, 2010, 2012) or Gries and Stoll (2009) develop a bottom-up approach called variability-based neighbor clustering to that problem, which can find temporal stages in diachronic data. This algorithm yielded interesting results, but is, in at least all applications I have seen, based solely on the exact data it is fed, meaning that it might be just as sensitive to corpus sampling artefacts/peculiarities as the case study in Section 1 and might, therefore, perhaps benefit from being complemented by the bootstrapping approach outlined here, which gives rise to stages via (lack of) overlap of the data ellipses resulting from the resampling. This idea should be explored further.

The present approach can also be improved. For instance, while the above identification of similar words to aid the interpretation of Zipf scale frequencies and ranges are already helpful, there is one really important way in which I would like to see frequency comparisons be improved. It is customary to compare frequencies of linguistic elements across different and especially differently sized corpora with normalized frequencies (e.g., pmw) or Zipf scale frequencies, but such comparisons are still potentially very misleading because, while they involve a correction for different corpus sizes, they do not correct for how the compositions of the corpora (or, here, corpus decades as a whole) have changed. For instance, a word type w_1 with a Zipf scale frequency z and the range value r in the 1960s decade might be the 1000th most frequent and the 2000th most evenly dispersed word type in that corpus decade, but a word type w_2 might have the same Zipf scale frequency z and the same range value r in another corpus decade, but, be the 500th most frequent and the 1000th most evenly dispersed word type simply because of how the word type distributions differ across the two corpus decades. Gries (2021b) uses this correction and shows that the results for the singular forms *gender* and *sex* (using ranks of frequencies and *DP*-values) support the results presented above and indicate that *gender* moves up considerably in the frequency and dispersion rankings (i.e., it becomes more frequent and evenly distributed than many other words per decade), whereas *sex* remains relatively the same across all decades. This kind of computation is computationally *extremely* demanding because it means that, even if one is only interested in two word types, one still needs to compute frequencies and dispersions of all word types in the corpus (parts) under consideration (for computing ranks for the words of interest) and one needs to do so once for every bootstrapping iteration. However, this strategy would allow for comparisons that are not just based on the values of two words, but take all the corpus changes into consideration. Hopefully, modern computing resources will make approaches such as these more feasible.

Regardless of the additional applications, extensions, and improvements we can come up with, I hope that the advantages of the current suggestions are attractive enough to make the field consider these suggestions and deviate from reporting just point estimates and/or simplistic chi-squared/G^2 applications in general corpus linguistics, theoretical applications (e.g., within cognitive or usage-based linguistics), or psycholinguistic experimentation, as well as in more applied fields such as learner corpus research or, like here, legal/forensic contexts. Hopefully, the increased reliability resulting from the above will allow the field to make further progress.

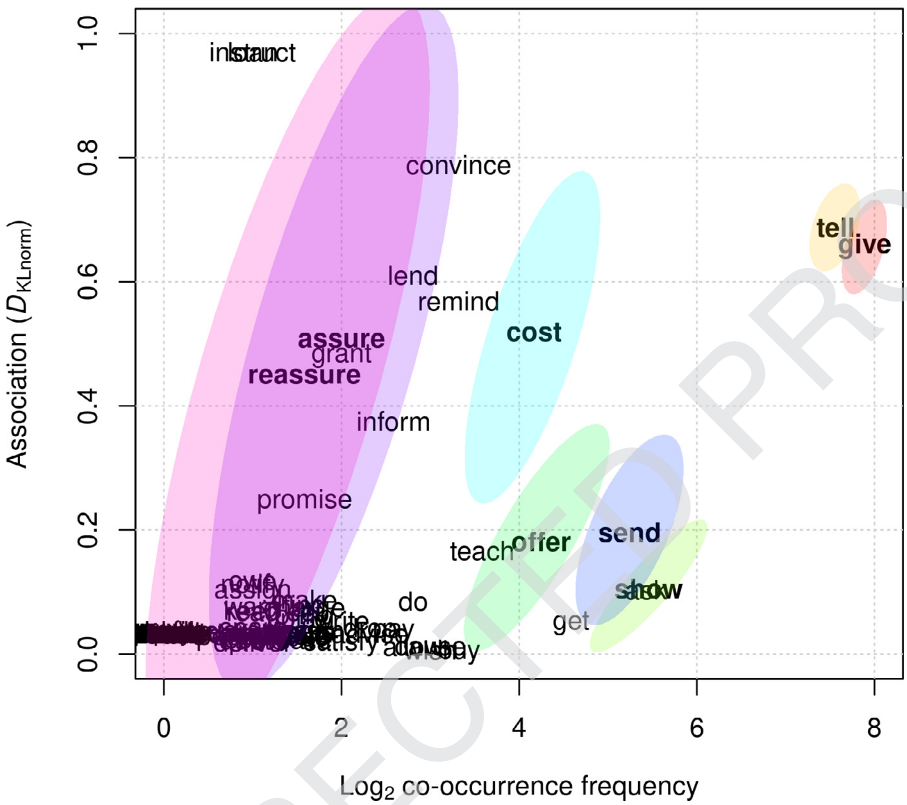

Fig. 9. The uncertainty of frequency/association for multiple lemmas.

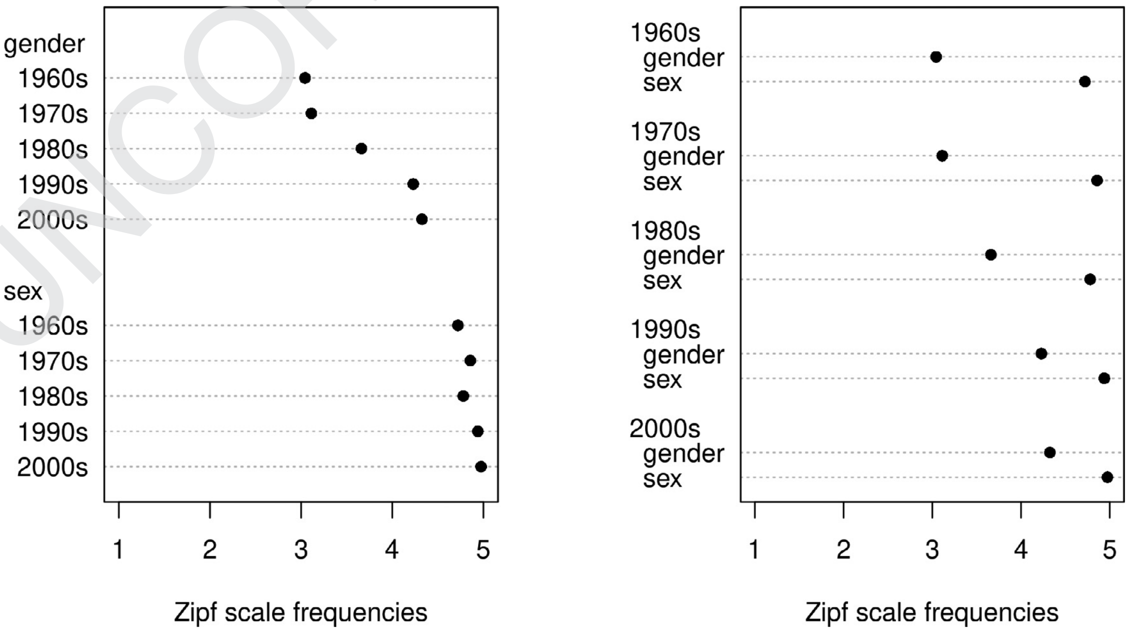

Fig. 10. Dotchart of point estimates of frequencies of gender/sex.

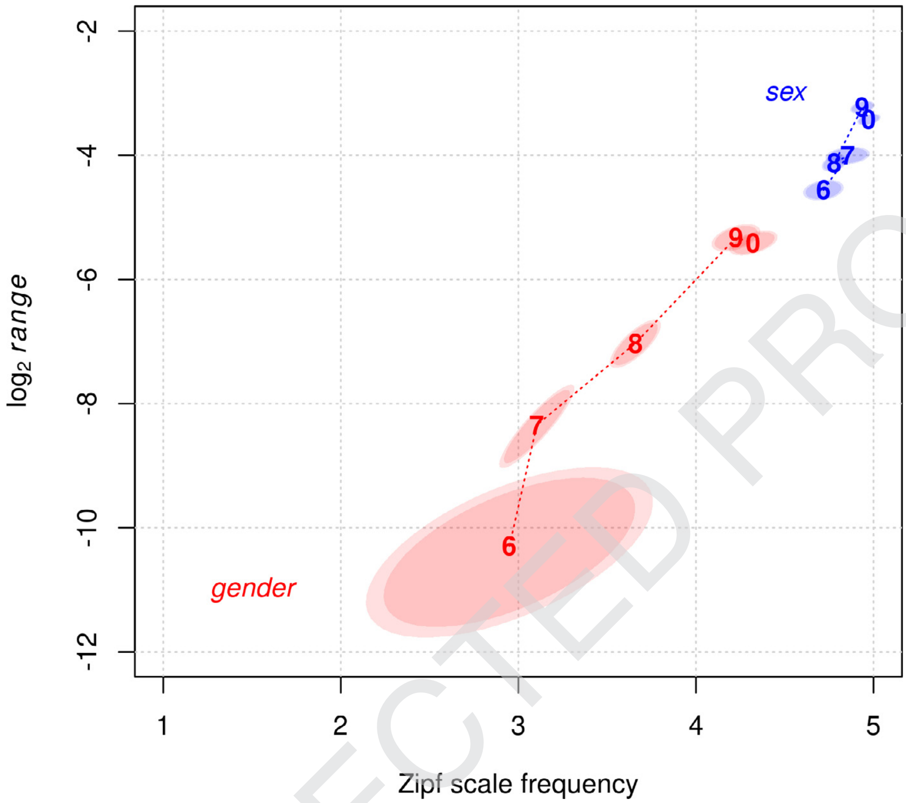

Fig. 11. Zipf scale frequencies against log ranges (bootstrapped).

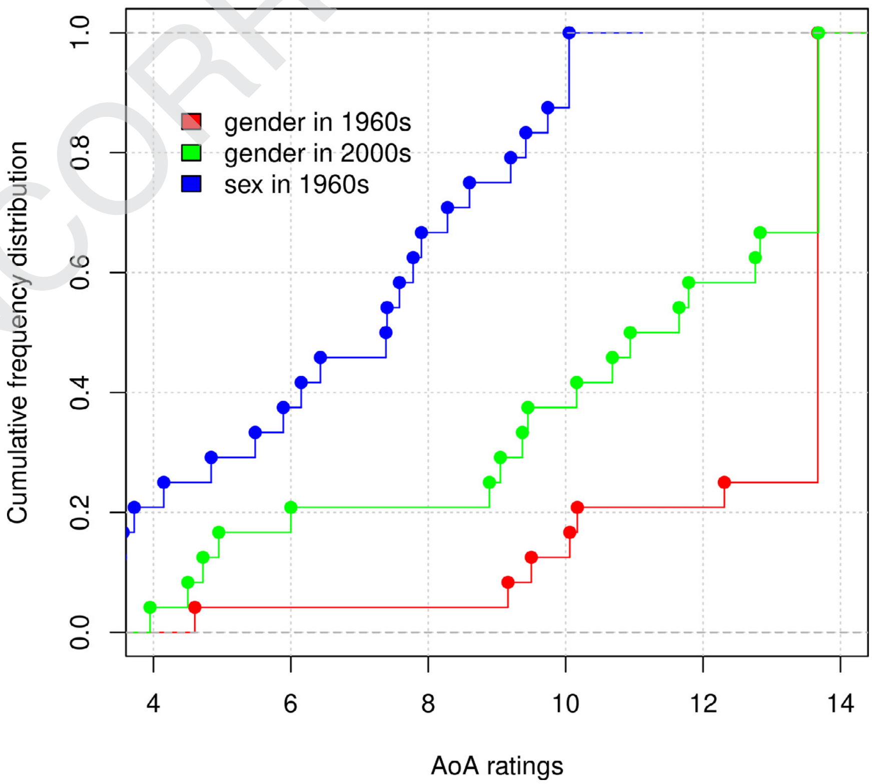

Fig. 12. AoA distributions for words similar to gender/sex in three COHA decades.

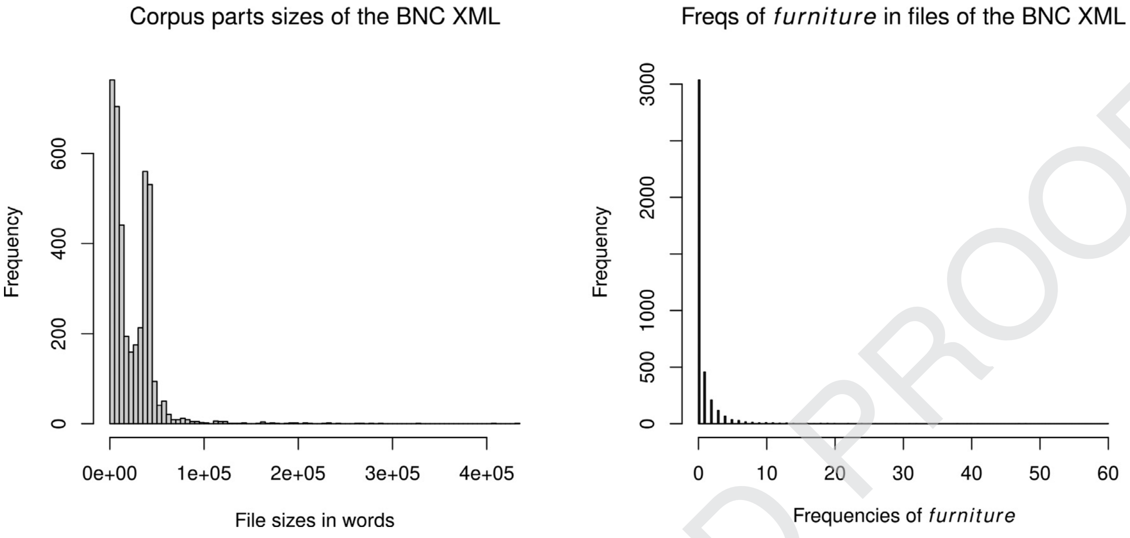

Fig. 13. Frequencies of corpus part sizes and of furniture in the BNC.

Declaration of Competing Interest

I declare 'no conflict of interest'.

References

Aijmer, Karin., Granger, Sylviane, Hung, Joseph, & Petch-Tyson, Stephanie (2005). Modality in advanced Swedish learners' written interlanguage. *Computer learner corpora, second language acquisition, and foreign language teaching* (pp. 55–76). Amsterdam & Philadelphia: John Benjamins.
Altenberg, Bengt., Granger, Sylviane, Hung, Joseph, & Petch-Tyson, Stephanie (2005). Using bilingual corpus evidence in learner corpus research. *Computer learner corpora, second language acquisition, and foreign language teaching* (pp. 37–54). Amsterdam & Philadelphia: John Benjamins.
Baayen, R. Harald (2011). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics, 11*(2), 295–328.
Biber, Douglas, & Egbert, Jesse (2018). *Register variation online*. Cambridge: Cambridge University Press.
Biber, Douglas, Reppen, Randi, Schnur, Erin, & Ghanem, Romy (2016). On the (non) utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics, 21*(4), 439–464.
Burch, Brent, Egbert, Jesse, & Biber, Douglas (2017). Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science, 3*, 189–216.
Church, Kenneth W. (2000). Empirical estimates of adaptation: The chance of two Noriegas is closer to $p/_2$ than p^2. In *Proceedings of the COLING 2000 (The 18th international conference on computational linguistics)* np.
Connor, Ulla, Precht, Kristen, Upton, Thomas, Granger, Sylviane, Hung, Joseph, & Petch-Tyson, Stephanie (2005). Business English: learner data from Belgium, Finland, and the US. *Computer learner corpora, second language acquisition, and foreign language teaching* (pp. 175–194). Amsterdam & Philadelphia: John Benjamins.
Culpeper, Jonathan, Demmen, Jane, Biber, Douglas, & Reppen, Randi (2015). Keywords. *The Cambridge handbook of English corpus linguistics* (pp. 90–105). Cambridge: Cambridge University Press.
Duyck, Wouter, Desmet, Timothy, Verbeke, Lieven P. C., & Brysbaert, Marc (2004). WordGen: A tool for word selection and non-word generation in Dutch, German, English, and French. *Behavior Research Methods, 36*(3), 488–499.
Egbert, Jesse, LaFlair, Geoffrey T., Phakiti, Aek, De Costa, Peter, Plonsky, Luke, & Starfield, Sue (2018). Statistics for categorical, nonparametric, and distribution-free data. *The palgrave handbook of applied linguistics research methodology* (pp. 523–539). London: Palgrave Macmillan.
Egbert, Jesse, Plonsky, Luke, Paquot, Magali, & Gries, Stefan Th. (2020). Bootstrapping techniques. *A practical handbook of corpus linguistics* (pp. 592–610). Berlin & New York: Springer.
Ellis, Nick C., Römer, Ute, & O'Donnell, Matthew Brook (2016). *Usage-based approaches to language acquisition and processing*. New York: Wiley-Blackwell.
Eskridge, William N., Slocum, Brian G., & Gries, Stefan Th. (2021). The meaning of sex: Dynamic words, novel applications, and original public meaning. *Michigan Law Review, 119*(7), 1503–1580.
Evert, Stefan. (2006). How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik, 54*(2), 177–190.
Fox, John, & Weisberg, Sanford (2019). *An {R} Companion to applied regression* (3rd ed.). Thousand Oaks CA: Sage.
Gardner, Dee, & Davies, Mark (2014). A new academic vocabulary list. *Applied Linguistic, 35*(3), 305–327.
Gilquin, Gaetanelle, Granger, Sylviane, Mukherjee, Joybrato, & Hundt, Marianne (2011). From EFL to ESL: evidence from the international corpus of learner English. *Exploring second-language varieties of English and learner englishes: bridging a paradigm gap* (pp. 55–78). Amsterdam & Philadelphia: John Benjamins.
Gries, Stefan Th (2006). Exploring variability within and between corpora: some methodological considerations. *Corpora, 1*(2), 109–151.
Gries, Stefan Th, Brdar, Mario, Gries, Stefan Th., & Žic Fuchs, Milena (2011). Corpus data in usage-based linguistics: What's the right degree of granularity for the analysis of argument structure constructions? *Cognitive linguistics: convergence and expansion* (pp. 237–256). Amsterdam & Philadelphia: John Benjamins.
Gries, S. Th. (2020). Analyzing dispersion. In Magali Paquot & Stefan Th. Gries(eds.), A practical handbook of corpus linguistics, 99-118. Berlin & New York: Springer.
Gries, S. Th (2021a). *Statistics for linguistics with R. 3rd rev. & ext. ed.* Boston & Berlin: De Gruyter.
Gries, S. Th (2021b). Corpus linguistics and the law: extending the field from a statistical perspective. *Brooklyn Law Review, 86*(2), 321–356.
Gries, S. Th. (2021c). What do (most of) our association measures measure (most)? Association? *Journal of Second Language Studies*.
Gries, S. Th. (2021d). What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies*.
Gries, St Th. & Durrant, P. (2020). Analyzing co-occurrence data. In MagaliPaquot & Stefan Th. Gries (eds.), A practical handbook of corpus linguistics, 141-159. Berlin & New York: Springer.
Gries, S. Th., Philip, Durrant, & Paquot, Magali (2021). Analyzing co-occurrence data. *A practical handbook of corpus linguistics* (pp. 141–159). Berlin & New York: Springer.
Gries, Stefan Th., & Hilpert, Martin (2008). The identification of stages in diachronic data: variability-based neighbor clustering. *Corpora, 3*(1), 59–81.

Gries, Stefan Th., & Stoll, Sabine (2009). Finding developmental groups in acquisition data: variability-based neighbor clustering. *Journal of Quantitative Linguistics, 16*(3), 217–242.
Gries, Stefan Th., & Hilpert, Martin (2010). Modeling diachronic change in the third person singular: a multi-factorial, verb- and author-specific exploratory approach. *English Language and Linguistics, 14*(3), 293–320.
Gries, Stefan Th., Hilpert, Martin, Nevalainen, Terttu, & Closs Traugott, Elizabeth (2012). Variability-based neighbor clustering: a bottom-up approach to periodization in historical linguistics. *The oxford handbook on the history of English* (pp. 134–144). Oxford: Oxford University Press.
Hofland, Knud, & Johansson, Stig (1982). *Word frequencies in British and American English*. Bergen, Norway: The Norwegian Computing Centre for the Humanities.
Kuperman, Victor, Stadthagen-Gonzalez, Hans, & Brysbaert, Marc (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods, 44*, 978–990.
LaFlair, Geoffrey T., Egbert, Jesse, Plonsky, Luke, & Plonsky, Luke (2015). A practical guide to bootstrapping descriptive statistics, correlations, t tests, and ANOVAs. *Advancing quantitative methods in second language research* (pp. 46–77). New York: Routledge.
Larson-Hall, Jennifer, & Herrington, Richard (2009). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics, 31*(3), 368–390.
Laufer, Batia, & Waldman, Tina (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), 647–672.
Leech, Geoffrey, & Fallon, Roger (1992). Computer corpora: What do they tell us about culture? *ICAME Journal, 16*, 29–50.
Lijffijt, Jefrey, Papapetrou, Panagiotis, Puolamäki, Kai, Mannila, Heikki, Gunopulos, Dimitrios, Hofmann, Thomas, Malerba, Donato, & Vazirgiannis, Michalis (2011). Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping. In *Proceedings of the ECML-PKDD 2011 – part II* (pp. 341–357). Berlin: Springer.
Lijffijt, Jefrey., Nevalainen, Terttu, Säily, Tanja, Papapetrou, Panagiotis, Puolamäki, Kai, & Mannila, Heikki (2016). Significance testing of word frequencies in corpora. *Literary and Linguistic Computing, 31*(2), 374–397.
Manning, Christopher, & Schuetze, Hinrich (1999). *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.
Neff van Aertselaer, JoAnne & Caroline Bunce (2012). The use of small corpora for tracing the development of academic literacies. In Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin, & Magali Paquot (eds.), *A taste for corpora: in honour of Sylviane Granger* (pp. 63-83). Amsterdam & Philadelphia: John Benjamins.
Plonsky, Luke, Egbert, Jesse, Geoffrey, T., & LaFlair (2015). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics, 36*(5), 591–610.
Rice, Sally, & Newman, John (2005). Inflectional islands. In *Paper presented at the International cognitive linguistics conference, Seoul, South Korea*.
Scott, Mike, & Tribble, Christopher (1996). *Textual patterns: key words and corpus analysis in language education*. Amsterdam & Philadelphia: John Benjamins.
Slocum, Brian G., Gries, Stefan Th., & Solan, Lawrence (2021). *Amicus brief with the U.S. Supreme court in the case Gerald Lynn Bostock v*. Clayton County, GA: On Writs of Certiorari to the United States Courts of Appeals for the 11th, 2nd, and 6th Circuits.
Stefanowitsch, Anatol, & Gries, Stefan Th. (2003). Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics, 8*(2), 209–243.
van Heuven, Walter, J. B., Mandera, Pawel, Keuleers, Emmanuel, & Brysbaert, Marc (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176–1190.