Corpus Linguistics: Quantitative Methods

STEFAN TH. GRIES

Introduction

Corpus linguistics is an inherently distributional discipline: ultimately, everything a corpus linguist does boils down to observed frequencies of some linguistic element(s) in some (parts of a) corpus or corpora either on its/their own or in the proximity of some other linguistic element(s) or discourse features (e.g., moves and lexically coherent units). Because of this, much of corpus linguistics is also inherently statistical. This chapter surveys quantitative methods in corpus linguistics. First, I discuss a few specifically corpus-linguistic statistics, before I turn to several examples of general statistical methods in corpus linguistics. I close with some desiderata for future research and/or developments in this area.

Corpus-Linguistic Statistics

The most basic corpus-linguistic statistic is frequency of occurrence, which usually comes in two kinds: token and type frequencies: **Token frequencies** state how often a certain word, lemma, construction, morpheme, etc. type is attested in a corpus, while **type frequencies** state how many different tokens (i.e., types) are attested in a corpus or in a slot around a word/of a construction. For instance, one might state that the word *enormous* occurs 37 times in the Brown corpus (a token frequency) or that there are 16 different words/strings preceding *enormous* (a type frequency), with *an* and *the* being the most frequent (occurring 15 and 5 times, respectively).

Token frequencies of words are a widely used proxy to the words' commonness and are assumed to be correlated with words' degree of entrenchment in speakers' mental lexicons. A useful way to report token frequencies and make them more comparable across differently large corpora involves normalizing them to "per million words" (pmw). Since the Brown corpus has approximately 1 million words, one might report the token frequency of *enormous* as 37 pmw. A version of this kind of normalization that is increasingly popular is the so-called Zipfscale (van Heuven et al. 2014), which is computed as shown in (1). Thus, assuming for simplicity's sake that the Brown corpus has exactly 1 million words, the Zipfscale frequency of *enormous* would then be 4.568202.

$$\text{Zipfscale} = 3 + \log_{10}(\textit{observed frequency per million words}) \tag{1}$$

Type frequencies, by contrast, are used to quantify semantic generality (i.e., the degree to which a word has a vague, or general, rather than a very specific, narrowly definable meaning) and productivity (i.e., the degree to which a construction is freely used with new/different words; for instance, if a new adjective was created, chances are the highly productive -ly suffix can be attached to it to make it an adverb). For example, a construction such as the ditransitive, with its rather specific meaning/function of "transfer," admits a much smaller type frequency of words into its verb slot than much more general constructions such as the active voice or the *going-to* future; thus, productivity is often approximated with the number of hapaxes (elements occurring only once) in a certain linguistic context.

The Encyclopedia of Applied Linguistics. Edited by Carol A. Chapelle.

DOI: 10.1002/9781405198431.wbeal20003

Table 1 Frequencies of a word w with and without another word x

	Word x	*Words other than* x	*Sum*
Word *w*	*a*	*b*	$a + b$
Words other than *w*	*c*	*d*	$c + d$
Sum	$a + c$	$b + d$	$a + b + c + d = N$

While token frequency is a widely used proxy to an element's commonness, it can be very misleading, a fact not sufficiently recognized especially in psycholinguistic work. This is because frequencies alone cannot distinguish between, for instance, two words with equal frequencies but extremely different dispersions. **Dispersion** is a quantity that reflects how (un)evenly an element is distributed in a corpus. For example, the word *staining* has the same frequency as *enormous* in the Brown corpus, but all 37 instances of *staining* are in just one of the 500 parts of the Brown corpus whereas the 37 instances of *enormous* are distributed over 36 corpus parts. Similarly, most speakers/scholars would probably agree that the words *nothing* or *whether* are more basic, more widespread, learned earlier, processed faster, etc. than the word *council*, yet in the British National Corpus, these words' frequencies are virtually identical—it is only their dispersion that confirms everyone's intuitions.

Many dispersion measures are available. One of the most widely used, but also the least informative one, is the range, which indicates how many parts of a corpus contain the element in question. For example, the range of *enormous* in the Brown corpus is $^{36}/_{500}$—but this ignores both the sizes of the corpus parts and the numbers of occurrences in each. Another frequent measure is Juilland's *D*, but recent comparative studies by Gries (2010, 2022b), Biber et al. (2016), and Burch et al. (2017) have shown that it is actually somewhat problematic and that Gries's *DP* is superior at least in their applications. *DP* is easy to compute:

1. convert the frequencies with which a word occurs in each corpus part into proportions (i.e., proportions of the overall frequency of the word);
2. convert the corpus part sizes into proportions (of the overall corpus size);
3. compute the pairwise differences of these proportions, take their absolute values, sum them up, and divide by 2.

This value falls into the interval (0, 1), with higher or lower values indicating a clumpier or more even distribution of an element in a corpus respectively. *DP* for *staining* in the Brown corpus is 0.998 (reflecting its clumpiness) while *DP* for *the* is 0.094 (reflecting its omnipresence).

The final main corpus-linguistic statistic quantifies association (and keyness). **Association** refers to the degree to which the presence of one element affects the probability of occurrence of another element. The three most common types of association are

- **collocation**, which refers to words co-occurring with other words;
- **collostruction**, which refers to words co-occurring with syntactic constructions;
- **keyness**, which refers to words preferentially occurring in one corpus as opposed to another.

In nearly all cases, association is quantified on the basis of 2 × 2 co-occurrence tables as the one schematically exemplified in Table 1, from which many different statistics can be computed. The cell frequency *a* represents the number of times *w* and *x* co-occur in the corpus, the frequency $a + b$ is the frequency of *w* in the corpus, the frequency $a + c$ is the frequency of *x* in the corpus, and *N* is the corpus size.

The most widely used association measures are probably the log-likelihood value G^2 and pointwise Mutual Information *MI*, which in the case of Table 1 would indicate the strength of the mutual attraction/repulsion of *w* and *x* in the corpus (see Evert 2009). By now, the field is also seeing a

growing interest in directional association measures such as ΔP, which can distinguish the attraction of w to x from that of x to w (see Gries, 2013).

The next section turns to the use of general statistical methods in corpus linguistics.

General Statistics in Corpus Linguistics

In general, it is hard to think of a general statistical technique that would not be applicable to corpus data at all. Expositorily, it makes sense to distinguish supervised and unsupervised methods. The former usually involve a response variable with known outcomes (e.g., a phonological, lexical, or constructional choice or the presence/absence of some element) and the analyst uses predictive modeling techniques to determine whether one or more predictors are likely responsible for how/whether the response variable is realized. The simplest such methods essentially reduce to traditional statistics such as a chi-squared test or a bivariate correlation (such as Pearson's r), but in nearly all quantitative corpus studies, more powerful techniques are required. The unsupervised methods usually involve many different variables without necessarily a specific response variable and the analyst uses exploratory methods to identify potentially interesting correlational patterns in the data.

Supervised Methods/Predictive Modeling

Regression Modeling

A particularly frequent kind of statistical application in corpus linguistics is some form of binary logistic regression modeling, that is, a regression modeling approach that involves a binary response variable that is predicted on the basis of one or more linguistic and extralinguistic predictors that are hypothesized to be correlated with speakers' choices. (In cases where one has only one categorical predictor, such a regression is very similar to a chi-squared test.) Generally, it is probably fair to assume that the vast majority of statistical analyses in corpus linguistics involve categorical response variables: the presence or absence of one or more linguistic elements. Numeric response variables do occur in corpus work (e.g., durations of words or formant frequencies of vowels in spoken corpora) but are much rarer. The main advantages of such regression models are that (a) several predictors can be evaluated at the same time, which is more useful than multiple single-predictor analyses; and (b) interactions between predictors can be included, which means one can determine whether some predictors strengthen, weaken, or even annul the effects of others. Of particular interest in recent years is the application of generalized linear mixed-effects models, that is, regression models that can accommodate random effects to deal with repeated measurements (when there are multiple data points by a speaker or for a certain lexical item) and with hierarchical corpus structure; see Barth & Kapatsinski (2018) and Gries (2021, chap. 6) for overviews.

Tree-Based Methods

Another growing family of predictive modeling methods involves tree-based approaches such as classification and regression trees and/or (random) forests. Trees are based on trying to recursively bifurcate the data into two parts such that the response variable is predicted as well as possible; random forests add two layers of randomness to this process, which can help reduce the impact of a variety of problems that can make regression modeling of observational data very difficult. These problems include (a) overfitting (the fact that models might not generalize well to unseen data), (b) collinearity (the fact that predictors may be highly correlated, which makes interpreting results different), (c) data sparsity (unlike regression models, random forests in particular can handle smaller data sets with many predictors), and others. Random forests are also attractive because

they perform very well on many predictive tasks often outperforming regression models and simple trees (see Levshina, 2020; Gries, 2021, chap. 7 for recent overviews).

Unsupervised/Exploratory Methods

A frequent exploratory method is **hierarchical clustering**, a method where an algorithm (a) determines how similar each case in one's data is to each other case using a similarity method defined by the user and then (b) groups together cases into clusters that have a high degree of within-cluster similarity and a low degree of between-cluster similarity. This bottom-up approach can be very useful to see how one's seemingly very diverse data set actually exhibits interpretively useful groups with rather distinct characteristics. For instance, one can use cluster analyses to study the systematic co-variation of phonetic features in corpus data, to identify semantic classes instantiated by verbs that fill certain constructional slots, or to determine chronological stages in diachronic data. Another form of clustering is ***k*-means clustering** where the user provides an algorithm with data, but then also a number of clusters k that the algorithm is supposed to detect in the data. For instance, an analyst might suspect that data exhibit a tripartite structure and set k to 3 to have the analysis of three groups (see Moisl (2015) for an introduction to cluster analysis).

Another powerful tool is (multiple) **correspondence analysis**, an exploratory method that is conceptually related to principal components analysis and that helps identify structure in multidimensional frequency data. Desagulier (2017, Section 10.4-5) discusses how this method has been used to identify instructive patterns in the frequency distribution of clause-final pragmatic particles across different social classes and in different locations in the United Kingdom.

While both kinds of exploratory methods discussed here require several sometimes technical analytical decisions, they are powerful tools to complement analysts' intuitions for data sets whose size and complexity rule out a mere eye-balling of the data.

Desiderata

While statistical analyses in corpus linguistics have much improved recently, much remains to be done, learned, and developed. Three areas deserve particular mention. First, the majority of studies still ignore the **multi-level structure of corpus data**. While mixed-effects modeling now more commonly addresses the problems of repeated measurements, too many studies still use the most basic statistical methods (chi-squared tests, simple correlations and regressions, etc.) and many studies still ignore the hierarchical structure of corpora, that is, the fact that, typically, speakers are nested into files/documents, which are nested into subregisters and/or registers. This can have a profound influence on analyses because variability that is not correctly attributed to speakers or registers might be incorrectly attributed to linguistic predictors such as animacy, length, and definiteness, anticonservatively exaggerating their perceived influence (see Gries, 2018 for discussion of how chi-squared/log-likelihood tests usually must be replaced with mixed-effects models).

A second problem that is only very slowly being recognized is that **many corpus-linguistic statistics conflate multiple levels of correlated information**. For example, most dispersion measures are computed such that they are extremely correlated with frequency (rather than measuring dispersion as a conceptually distinct dimension). The same is true of many association/keyness measures, and quantities such as adjusted frequency even try to sell this as an advantage. However, this kind of conflation leads to a huge amount of information loss and corresponding validity problems. Corpus linguistics needs to ensure that dimensions of information that are conceptually distinct are measured separately and reliable than they currently are (see Gries, 2022b, for an example).

Finally, much corpus-linguistic work does too little in terms of quantifying the **uncertainty** that the corpus data come with. For example, studies of frequencies, dispersion, or association usually

do not come with much quantification of how variable the data are, and parametric confidence intervals are often not even applicable to corpus data. Given how many corpus studies rely on rankings (of associations, of keywords, etc.) and given how volatile corpus data can be, quantifying uncertainty of our results becomes essential, and recent work (e.g., Egbert & LaFlair, 2018; Gries, 2022a) exemplifies how bootstrapping can help. Bootstrapping is a resampling method that is based on (a) repeatedly drawing random samples (e.g., 500) from one's data (to simulate how one's data could have looked like if one had studied a different sample from a similar corpus) and (b) computing one's statistic on each of the 500 random samples to get an impression of how volatile the results really are. More and faster steps in these directions are bound to help corpus linguistics keep up with statistical methods that are already much more widely adopted in many other social sciences.

SEE ALSO: Collostructional Methods; Multiple Regression; Probability and Hypothesis Testing; Testing Independent Relationships

References

Barth, D., & Kapatsinski, V. (2018). Evaluating logistic mixed-effects models of corpus-linguistic data in light of lexical diffusion. In D. Speelman, K. Heylen, & D. Geeraerts (Eds.), *Mixed-effects regression models in linguistics* (pp. 99–116). Springer.

Biber, D., Reppen, R., Schnur, E., & Ghanem, R. (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, B., Egbert, J., & Biber, D. (2017). Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science*, *3*(2), 189–216.

Egbert, J., & LaFlair, G. T. (2018). Statistics for categorical, nonparametric, and distribution-free data. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 523–539). Palgrave Macmillan.

Evert, S. (2009). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (vol. 2, pp. 1212–1248). de Gruyter.

Gries, S. T. (2010). Dispersions and adjusted frequencies in corpora: further explorations. In S. T. Gries, S. Wulff, & M. Davies (Eds.), *Corpus linguistic applications: Current studies, new directions* (pp. 197–212). Rodopi.

Gries, S. T. (2013). 50-something years of work on collocations: What is or should be next …. *International Journal of Corpus Linguistics*, *18*(1), 137–165.

Gries, S. T. (2018). On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies*, *1*(2), 276–308.

Gries, S. T. (2021). Statistics for linguistics with R: A practical introduction. *3rd rev. &* (ext. ed.). Mouton de Gruyter.

Gries, S. T. (2022a). Towards more careful corpus statistics: Uncertainty estimates for frequencies, dispersions, association measures, and more. *Research Methods in Applied Linguistics*, *1*(1), 100002.

Gries, S. T. (2022b). What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies*, *5*(2).

Levshina, N. (2020). Conditional inference trees and random forests. In M. Paquot & S. T. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 611–643). Springer.

Moisl, H. (2015). *Cluster analysis for corpus linguistics*. John Benjamins.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.

Suggested Readings

Desagulier, G. (2017). *Corpus linguistics and statistics with R*. Springer.

Larson-Hall, J., & Herrington, R. (2009). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics*, *31*(3), 368–390.
Paquot, M., Gries, S., & Th. (Eds.). (2021). *Practical handbook of corpus linguistics*. Springer.