3

Quantitative Designs and Statistical Techniques

Stefan Th. Gries

3.1 Introduction

As is well-known, corpus linguistics is an inherently distributional discipline: Corpora really only contain strings of elements – letters/characters in the typical case of corpora as text files, phonemes or gestures in the growing segment of auditory or multimodal corpora – and analysts can determine their frequency of occurrence, frequency of co-occurrence, or their dispersion/distribution in corpora. Crucially, because of that, analysts must make careful choices regarding the quantitative design of their study because anything they are interested in – acquisition, language processing and change, meaning, communicative function/intention, speaker proficiency, and so on – typically requires two levels of quantitative design choices, which will also provide a structure to this overview.

At the first level of quantitative design, analysts need to consider *what specifically corpus linguistic statistics they need to compute* (or what versions of such statistics) to most meaningfully describe/explore their questions and/or test their hypotheses, and such considerations will need to be based on theoretical and linguistic issues as well as characteristics of the corpus linguistic statistics to be used. For a fairly straightforward heuristic classification, some corpus linguistic statistics involve little to no contextual information (e.g., frequencies as well as some keyness applications and dispersions), some involve limited contextual information (e.g., association measures for collocations, collostructions, and colligations), and some involve a lot of contextual information (e.g., concordance lines annotated for many contextual features).

However, such considerations are only the first step, because they usually lead to a whole host of other important quantitative design questions: For example, in a learner corpus study involving one native speaker corpus and two non-native/learner corpora, does one's hypothesis require some quantification of lexical complexity, diversity, or sophistication? If lexical sophistication is what is required,

- is it computed for all word types attested in a learner corpus or only word types exceeding a certain frequency and/or dispersion threshold? (And this even glosses over the question of what a token/type is, an important question to be revisited below.)
- what measure of lexical sophistication is computed – one that is only based on frequency of occurrence, one that is only based on dispersion, one that uses both, or even more dimensions of information?
- for both frequency and dispersion, what exact information is entered into the relevant formula(e)? For example, how does one measure frequency such that different corpus sizes are controlled for? Does one's dispersion measure include the information how often each word type is attested in each native and non-native text? Does it include the size of each text? Or does one choose a dispersion measure that does not even consider corpus parts but treats each corpus as one whole chunk?

As should be obvious by now, even for a superficially simple-seeming comparison of lexical sophistication scores across corpora representing three speaker populations, an increasingly multidimensional search space of initial choices has to be considered, but even after those choices have been made, the complexity does not end there: At a second level of quantitative design choices, analysts need to also consider *what statistical method(s) to use to describe and evaluate the end product of all design decisions at the first level*, which adds another set of questions to be answered; the following is a selection of some such questions roughly ordered by increasing generality:

- On what information level does one treat the values of whatever version of lexical sophistication one computed – ordinal? numeric?
- On what level of resolution does one compare the native and learner varieties? Does one use measures of central tendency of lexical sophistication measures for each corpus such as medians or means (or does one need some other more robust measure; see, e.g., Wilcox, 2017) normalized against some measure of statistical dispersion such as interquartile ranges or standard deviations (or some other more robust measure)? Or does one use a test sensitive to an overall shape of distribution?
- More abstractly, do the data come with a multilevel structure one needs to account for (as when many learner essays cover only a small range of topics for which we then have 'repeated measurements')?
- Even more abstractly, is one's goal description (answering the question of what is happening), inference (answering the question of what non-random patterns exist), prediction (the ability to generalize well to unseen data), attribution (the identification of causal pathways), or some combination of the above, each of which may require different statistical approaches?

In sum, because of the distributional nature of corpus data, any quantitative corpus study needs to make many interrelated design choices at these two levels that are hopefully motivated well by the linguistic research question(s),

correct or at least useful, and compatible with each other. This high number of quantitative researcher degrees of freedom is why knowledge of the discipline involving the analysis of frequencies/distributions – aka statistics – should obviously form a central component of corpus linguists' methodological knowledge. However, even compared to other related social sciences (e.g., psychology, communication, sociology, anthropology, among others) or other branches of linguistics (e.g., psycholinguistics, phonetics, sociolinguistics, etc.), much of corpus linguistics has begun to develop this methodological awareness only with some delay. Thus, corpus linguistics needs to 'catch up' on both these levels: (i) on the first level of statistics that are more directly related to central corpus linguistic tools such as frequency lists, association phenomena like collocations and collostructions, and dispersion; and (ii) on the second level of general statistics that are applied to the statistics from the first level.

In this overview, I will discuss statistical tools in corpus linguistics. Section 3.2 is devoted to the 'first level', that is, statistics directly involving corpus linguistic tools; Section 3.3 then turns to the 'second level', that is, statistics that are usually applied to concordances and their annotation. In each section and subsection, I will first discuss some commonly used methods and then provide some pointers to methods that are currently underutilized and whose exploration or wider use would most likely benefit the field. Section 3.4 will conclude with some more general comments.

3.2 Statistics on Core Corpus Linguistic Methods

In this section, I will be concerned with statistical methods that involve, or apply 'directly' to, the methods of frequency lists, collocations, and dispersion.

3.2.1 Frequencies of Occurrence

3.2.1.1 Frequency Lists

Frequencies of occurrence are obviously the most basic statistic one can provide for any linguistic phenomenon. They can come as either token or type frequencies and typically in one of the following three forms:

- raw frequencies: the frequency of *give* in the spoken component of the ICE-GB is 297;
- normalized frequencies: the frequency ptw (per thousand words) of *give*'s in the spoken component of the ICE-GB is ≈0.46575 or ≈465.75 pmw (per million words);
- logged frequencies: the (natural) log of the frequency of *give* in the spoken component of the ICE-GB is 5.693732.

Raw frequencies are easiest to interpret within one corpus, normalized frequencies are most useful when frequencies from differently sized corpora must be compared, and logged frequencies are useful because many psycholinguistic manifestations of frequency effects operate on a log scale and because testing for Zipfian distributions involves plots on a log-log scale. A maybe particularly useful way of reporting frequencies is one version of the Zipf scale (van Heuven et al., 2014), which is computed as shown in eq. (3.1). This measure has the advantages that it is already normalized and log-scaled and that it considers unobserved words in the calculation (much like smoothing methods would):

$$Zipfscale = 3 + log_{10}\frac{10^6 \cdot (obs.freq. + 1)}{corpussize_{tokens} + corpussize_{types}} \tag{3.1}$$

Most often, the frequencies that are reported are word frequencies in (parts of) corpora but many studies are also concerned with frequencies of morphemes, grammatical constructions, words in constructions, or n-grams/lexical bundles. Examples abound in

- learner corpus research, to document potential over-/under-use by learners compared to native speakers;
- language acquisition corpora, to document children's increasing type frequencies of, say, which verbs may fill the slot in a particular construction;
- historical linguistics, to document the increase/decrease of use of particular words or constructions over time.

In spite of the apparent straightforwardness of the above, there are still several underutilized methods and desiderata. One is concerned with what frequency distributions especially of elements in slots can reveal if one studies their (normalized) entropy as defined in eq. (3.2).

$$\text{a.} \quad H = \sum_{i=1}^{n} p(x) \cdot -log_2 p(x), \text{ with } log_2 0 := 0$$

$$\text{b.} \quad H_{norm} = {}^{H}\!/_{H_{max}} = {}^{H}\!/_{log_2 n} \tag{3.2}$$

The (normalized) entropy of a frequency distribution indicates how the frequencies are distributed. For example, two verb lemmas v_1 and v_2 might exhibit very different frequencies of their different morphological forms. Consider Figure 3.1, whose left panel represents the frequencies of the verb lemma GIVE in the ICE-GB and one can see that, while the forms are certainly not equally frequent, each one occurs at least somewhat frequently. That means, if one had to guess which verb form a single randomly chosen verb form of GIVE from the corpus would be, one would be wrong a lot of the time: Even the most preferred form (*give*) is only 4.2 times more frequent than the least preferred form (*gives*). However, the right panel represents analogous data for SING and here we have a more uneven/spiky distribution and the most preferred form (*singing*) is 22.5 times more frequent than the least preferred form (*sings* or *sung*), and this is what (normalized) entropy helps

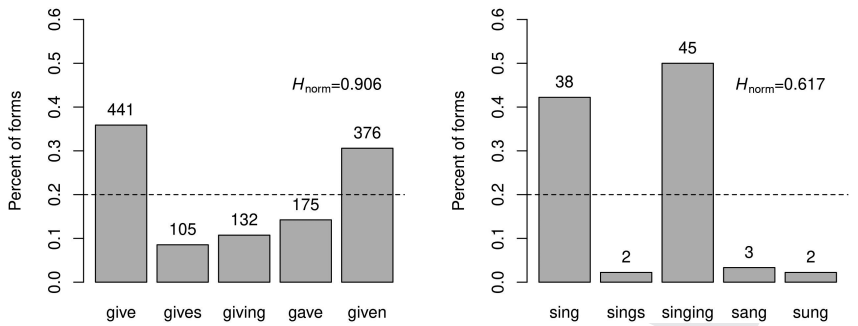

Figure 3.1 Normalized entropies (H_{norm}) of the forms of GIVE and SING in the ICE-GB.

quantify. Especially in studies concerned with the learnability of what can and cannot, or likes and does not like to, go in constructional slots, entropy values are a useful summary statistic complementing simple frequency distributions.

Even more interesting for frequency lists of words or n-grams is the question of what the counted units, or tokens, are or should ideally be – a question that is of course relevant far beyond the current context of frequency lists. This means researchers need to decide on what basis to tokenize: on the basis of

- existing part-of-speech annotation, which of course may mean that, for the sake of speed but also replicability or comparability with other studies using the same corpus, one might place a certain amount of trust in the theoretical foundations and reliability of the used tagging information;
- certain characters and their contexts (spaces, punctuation marks, etc.), which often also raises the question of how to deal with, say, hyphenated expressions and apostrophes.

One example of how tricky especially the former issue can become involves multi-word unit (MWU) tags in the British National Corpus XML, which identify units like *of course*, *such as*, *in terms of*, *as well as*, and even longer ones like *in so far as* or *a hell of a lot*, which are useful, but (i) many other useful ones are not included (e.g., *on the (one|other) hand* or *be that as it may*) and (ii) many MWUs that are included are not useful at all (e.g., *in vivo*. [with the period], *because er* [the disfluency *uh*] *of* or *on top top of*). The latter issue, on the other hand, arises with basically any kind of unannotated corpus (e.g., web-scraped and/or OCR-ed data) or annotated corpora whose annotation is not appropriate for, or compatible with, one's current research question. For the complex question of MWU tags again, many different strategies have been proposed in both corpus and computational linguistics but one crucial problem is that the length of the continuous or even discontinuous n-grams or lexical items needs to be variable so that one might find differently long elements such as *according to*, *in spite of*, *on the other hand*, *be that as it may*, and *the fact of the matter is*. There is corpus linguistic work on this (e.g., Kita et al., 1994; Daudaravičius & Marcinkevičienė, 2004; Gries & Mukherjee, 2010;

Brook & O'Donnell, 2011; Gries, 2022a, Ben Youssef, 2024) and much work in computational linguistics has been devoted to multi-word identification and discovery (e.g., early studies such as Nagao & Mori, 1994; or Ikehara, Shirai, & Uchino, 1996 and more recent work like McCauley & Christiansen, 2019, or the Python Partitioner project of Williams & Reagan, 2017), but the exact combination and weightings of features yielding good precision and recall – which may also be very different for different purposes – are still proving elusive. Given the importance of this for something as fundamental as tokenization, much more exploration and testing of MWU identification is required.

3.2.2 Keywords

A widespread application of frequency lists is the comparison of frequency lists of two corpora, often one (larger) reference corpus *R* and one (smaller and/or more specialized) target corpus *T* to determine which words are 'key' for *T* compared to *R*. One of the earliest approaches is Damerau's (1993) relative frequency ratio, but most keywords studies have used the association measure G^2 (see Dunning, 1993) instead. Consider Table 3.1 for the keyness computation for the word *college* in the Clinton–Trump corpus of presidential election speeches.

Computing G^2 for these observed data would require computing the expected frequencies (as one might do for a chi-squared test for independence) and then inserting observed and expected values into the formula in eq. (3.3), which returns a value that indicates that the word *college* is very key for Clinton's speeches:

$$G^2 = 2 \cdot \sum_{i=1}^{4} obs \cdot ln \frac{obs}{exp} = 213.7825 \tag{3.3}$$

Most approaches to keyness, however, suffer from the problem that they do not consider dispersion (see Section 3.2.3), that is, how evenly or unevenly a potential keyword is distributed across the target corpus, which is one reason why other measures have been proposed. Egbert and Biber (2019) develop a measure that replaces the observed frequencies of occurrence in tables like Table 3.1 with the numbers of parts the word in question does or does not occur in. Paquot and Bestgen (2009), by contrast, do not replace frequency with dispersion but suggest the use of a *t*-test that compares relative frequencies of a potential keyword in the parts of a corpus, which combines frequency

Table 3.1 *The keyness of* college *for Clinton (compared to Trump)*

	T (Clinton)	*R* (Trump)	Sum
Word: *college*	106	26	132
Other words	117,183	445,704	562,887
Sum	117,289	445,730	563,019

and dispersion in the computation. Finally, Gries (2024a) proposes to consider keyness a three-dimensional construct of frequency, dispersion, and association so that analysts can distinguish keywords that are key for different statistical reasons; he shows that, for example, words that are key for academic writing based on association are very different from words that are key for academic writing based on dispersion.

3.2.3 Frequencies of Co-occurrence

For many linguistic questions, frequency of occurrence of phenomenon *P* alone is not sufficient – rather, what is required is the frequency of *P* co-occurring with something else, say *S, T,* …. Typically, when *P*, *S*, *T*, … are words, this is referred to as *collocation* (and *P, S, T*, … are *collocates*); when *P* is a construction/pattern, this is referred to as *colligation* or *collostruction* (and *S*, *T*, … are called *collexemes* of *P*). In both cases, a central concern is being able to rank collocates/collexemes *S*, *T*, … in terms of their direction and strength of association with *P* (is the relation between *P* and *S*, *T*, … one of attraction or one of repulsion, and how strong is it?). More than 80 different measures have been discussed in the last few decades (see Pecina, 2010) but nearly all the most widely used ones derive from a 2 × 2 co-occurrence table such as Table 3.2, which is in fact conceptually virtually identical to Table 3.1 and provides the frequencies of co-occurrence of the verb form *give* and the ditransitive construction in the ICE-GB.

Different association measures have different statistical properties and can lead to very different results and rankings. As for keyness, G^2 is actually one of the most widely used ones and as a significance-based measure, it is very highly correlated with the observed frequency of co-occurrence. On the other hand, effect size measures such as the log odds ratio or pointwise mutual information (PMI) are much less dependent on co-occurrence frequency and reflect association more 'cleanly'; these three measures are exemplified in eq. (3.4).

$$\begin{aligned}
&\text{a.} \quad G^2 = 2 \cdot \sum_{i=1}^{4} obs \cdot ln\tfrac{obs}{exp} = 1451.9307 \\
&\text{b.} \quad log\ odds\ ratio = ln\tfrac{234¼209}{1586¼136878} = 4.57086 \\
&\text{c.} \quad Pointwise\ Mutual\ Information = log_2 \tfrac{234}{exp.\text{co-occ.}=5.804315} = 5.33324
\end{aligned} \tag{3.4}$$

Table 3.2 *Schematic co-occurrence table of token frequencies for association measures*

	Construction: *ditransitive*	Other constructions	Sum
Verb form: *give*	234	209	443
Other verb forms	1,586	136,878	138,464
Sum	1,820	137,087	138,907

Because of their different statistical behaviors and different applications, it is not straightforward to single out one association measure as 'the best' and the field has still much to explore. Two areas are particularly noteworthy. The first of these is only concerned with collocations and is concerned with the range of words around a word *P* that are included. Just like with n-grams, practitioners usually seem to make an arbitrary choice, and frequent choices are 4, 5, or 10 words to the left and to the right, yielding context windows of 8, 10, or 20 words. However, Mason (1997, 1999) provides a much better solution to this problem, which is unfortunately hardly ever used. He proposes to include larger context windows and then for each slot before or after *P* he computes the entropy of the frequency distribution of the collocates in that slot, which allows analysts to decide on what windows sizes to choose and what collocate slots to investigate in a data-driven fashion.

The second area in need of additional research is concerned with the association measures per se. While the number of measures is large, just about all – and all that are regularly used – have two potentially undesirable characteristics:

- They are bidirectional, which means they assign one value to, say, two collocates P and S and do not distinguish whether perhaps the association P→S is smaller than that of S→P.
- They are based on token frequencies of *P* and *S* alone and do not take into account the type frequencies of words co-occurring with *P* and *S* (let alone the entropies of these type frequencies; see Gries, 2012).

There are three main unidirectional measures that have been proposed, but we need much more exploration of them and maybe others especially because – coming back to the discussion of quantitative designs at the beginning – many phenomena one might be interested in may involve unidirectional association, which means that, to do justice to the research question, we have to choose our measures accordingly. Ellis (2007) discusses the measure ΔP from the associative learning literature in corpus linguistics, which was then used in Ellis and Ferreira-Junior (2009) as a measure of verb-to-construction and construction-to-verb association and was very similar result-wise to $p_{\text{Fisher-Yates exact test}}$; Gries (2013) finds that, when ΔP is applied to multi-word units and 2-grams, it can identify different degrees and kinds of directional association quite well. Michelbacher, Evert, and Schütze (2011) also discuss different approaches – conditional probabilities and ranks of association measures – and while the latter are promising, the computational effort they involve is very large. Finally, Gries (2024a) proposes to use both directions of the Kullback–Leibler divergence (KLD) as a measure that is similar to the widely used G^2 but directional and less correlated with co-occurrence frequency.

3.2.4 Dispersion

Another topic that is even more important but at least as understudied is the notion of dispersion, the degree to which any frequency of *P* – frequencies of occurrence or co-occurrence – is extremely sensitive to how evenly *P* is distributed in a corpus. Ever since some early work in the 1970s by Carroll, Juilland, Rosengren, and others (see Gries, 2008 for a comprehensive overview), researchers have attempted to develop measures of dispersion that indicate how (un)evenly an item may be distributed in a corpus. These can either be used on their own or they can be used to compute so-called adjusted frequencies, that is, word frequencies that are adjusted downwards proportionally to the degree that words are unevenly distributed. For instance, both *amnesia* and *properly* occur 51 times in the 500 files of the ICE-GB but not only is it obvious which of the two is more entrenched, acquired earlier, and more useful for foreign-language learners, this is also reflected in their dispersion: *amnesia* and *properly* occur in 2 and 47 files of the ICE-GB respectively so their ranges of $^{2}/_{500}$ and $^{47}/_{500}$ and Juilland's adjusted frequencies (so-called usage coefficients) are ≈14 and ≈43.5 respectively, which underscores what, here, is intuitively clear: *amnesia* is much more specialized.

There has been some recent research on dispersion. Gries (2008) proposed a new measure that performed well in his own validation (against reaction times) and in Biber et al.'s (2016) study, a study that also indicates that one of the most widely used dispersion measures – Juilland's *D* – does not perform as well as its wide use might suggest. Another interesting study is Burch, Egbert, and Biber (2017), who also propose and test a new dispersion *DA* in a simulation study and with 150 words from the BNC. Finally, Gries (2022b, 2024a) proposes again to use KLD as a dispersion measure but also discusses the fact that most dispersion measures are very highly correlated with frequency, which can be problematic because it means that many measures can only return low dispersion values for infrequent words. He discusses the example of *enormous* and *staining* in the Brown Corpus, which are both relatively infrequent in the corpus (occurring 37 times) and which score fairly small dispersion values on many measures even though they occur in 36 and 1 parts respectively. In other words, many dispersion measures fail to see that, *given its frequency*, *enormous* is nearly perfectly evenly dispersed; Gries (2022b, 2024a) discusses a solution for this (and shows how it improves the predictability of reaction times), but we need more research on this.

Given the generality of this problem – any statistic in corpus linguistics is ultimately based on frequencies in parts of corpora – both dispersion and the intimately related notion of corpus homogeneity should always be considered potential threats to our studies. Gries (2006) exemplifies how even the simplest of phenomena – frequencies of present perfects – can exhibit large variability on different levels of corpus granularity and how the degree to which, say, constituent order alternations can be accounted for can differ hugely between different corpus parts.

3.3 General Statistics

In this section, I will now turn to surveying statistical tools that are often applied to the first-level corpus statistics discussed in the previous section and/or to the annotation of concordance data, that is, to data that emerge from the description – linguistic, contextual, or otherwise – of concordance data; Section 3.3.1 is concerned with confirmatory statistics (and mentions descriptive statistics in passing); Section 3.3.2 with exploratory statistics.

3.3.1 Confirmatory/Hypothesis-Testing Statistics

Confirmatory statistics can be classified according to two main characteristics:

- the number of independent variables, or predictors: either the design is monofactorial (there is only one predictor; see Section 3.3.1.1) or multifactorial (there are two or more predictors; see Section 3.3.1.2);
- the nature of the dependent variable(s), or the response(s), which is usually either categorical (e.g., animacy of possessor: *animate* vs. *inanimate*) or numeric (e.g., lengths of recipients and patients in characters) and which, thus, affects the choice of statistic chosen.

3.3.1.1 Monofactorial Statistics

Monofactorial statistical analysis has been relatively frequent in corpus linguistics for quite a while. Aarts (1971) is a classic early case in point studying the distribution of NP types in English clauses. She conducts a variety of chi-squared tests to explore, for instance, what kinds of NPs occur in subject slots and finds that subject slots prefer structurally light NPs. Another well-known application of chi-squared tests is Leech and Fallon's (1992) study of what word frequency differences between the Brown and the LOB corpus might tell us about cultural differences between the US and the UK. They use a difference coefficient and chi-squared tests to identify words key for American or British English. As a final example, Mair et al. (2002) compare part-of-speech frequencies between the 1960s LOB corpus and its 1990s counterpart FLOB; using G^2 they find that frequencies of nouns increase considerably over time.

Turning to other monofactorial explorations, Schmitt and Redwood (2011) is an example for the use of correlations. They use the Pearson product-moment correlation r to address the question of whether learners of English knowledge of phrasal verbs is related to the verbs' frequency (in the British National Corpus) and find a significant though noisy correlation; in addition, they use a t-test to see whether learners' performance differs between reception and production, and they do. Another example from the same domain is Durrant and Schmitt (2009), who compare the use of adjective-noun and noun-noun collocations by learners with that of native speakers, which were

extracted from essays and whose strength was quantified using the *t*-score and PMI. These values were classified into seven and eight bands respectively so the authors could explore NS and NNS' use collocations of particular strengths with *t*-tests. Results for the *t*-scores indicate that NNS make greater use of collocations in terms of tokens but not when type variability is considered as well, whereas results for *PMI* indicate that NNS make less use of collocations in terms of tokens when type variability is considered as well.

Two comments may already be pertinent here. One is that it is important that corpus linguists explore in detail whether the assumptions of the significance tests of their statistics are met. Significance tests of Pearson's *r* or the *t*-test require bivariate normality or normality and variance homogeneity respectively, but often there is no mention of whether these assumptions were in fact explored. With observational data, which are often Zipfian-distributed and sparse, alternatives such as Kendall's τ or the *U*-test (as in Borin's 2004 study of *n*-gram-type frequencies of native and non-native speakers) or the very much underused Kolmogorov–Smirnov test may often be more appropriate.

The other general point is that corpus linguists need to realize more that no linguistic phenomenon is monofactorial ever. Any monofactorial test can only be a (dangerous) shortcut, given that what is really required for confirmatory statistics is a kind of analysis that combines three characteristics (see Gries & Deshors, 2014; Gries 2024b):

- They are multifactorial in that they simultaneously consider multiple causes for linguistic choices.
- They involve interactions between level-1 predictors – predictors that are annotated at the same level as the linguistic choice being modeled – so that one can determine whether a particular predictor has the same effect regardless of other predictors' values.
- They involve interactions between level-1 predictors and level-2 variables (to, e.g., catch speaker-specific effects) or higher-level predictors like L1 (is the speaker a native speaker or a learner of some variety?), Register (which register/genre is a data point from?), Time (which time period is a data point from?), etc., because without such interactions it is impossible to determine whether the level-1 predictors have the same effect for all speakers and in each L1/variety, in each register, at each time period, etc.

Thankfully, studies taking this into consideration are (slowly but steadily) becoming the mainstream and the following section discusses this in more detail.

3.3.1.2 Multifactorial Statistics

One of the most important and most widespread tools in (predictive) modeling approaches in corpus linguistics is the generalized linear model (GLM) and its extensions, a family of regression models in which a dependent variable is

modeled as a function of one or more predictors. Crucially, in GLMs and its extensions, the dependent variable can be of different kinds: They can be

- numeric, where the GLM boils down to 'regular' linear regression models;
- ordinal, so one might compute an ordinal regression;
- binary or categorical, so one might compute a binary logistic regression (see above) or a multinomial regression;
- frequency counts, so one might compute a Poisson regression or a negative binomial regression.

In the same way, predictors can also be numeric, ordinal, binary or categorical variables, or any interactions between such variables, and the results of such regressions are predictions (either raw values or predicted probabilities of outcomes). The earliest such confirmatory studies that I am aware of – see below for earlier multivariate exploratory methods – are Leech, Francis, and Xu's (1994) use of loglinear analysis to explore the alternation between *of*- and *s*-genitives and Gries's (2000, 2003) use of linear discriminant analysis to study particle placement, the alternation of *John picked up the squirrel* and *John picked the squirrel up*. Following these studies and various replications and extensions – see Kendall, Bresnan, and van Herk (2011) on the dative alternation, Diessel and Tomasello (2005) on particle placement in child language acquisition, Hinrichs and Szmrecsanyi (2007) on genitives, and so on – such regression analyses have become adopted more frequently.

Most of these applications involve binary logistic regressions, but multinomial regression is also slowly becoming more mainstream. Han, Arppe, and Newman (2017) model the use of five Shanghainese topic markers on the basis of TopicLength, TopicSyntcat, Genre, and other variables. An example for ordinal logistic regression is Onnis and Thiessen (2012), who model levels of syntactic parse depths in English and Korean as a function of *n*-gram frequencies and two conditional probabilities and show, for example, that cohesive phrases tend to be more frequent and that "the patterns of probability that support syntactic parsing are clearly reversed in the two languages."

The more widespread adoption of these tools is already huge progress, and even greater progress is the increasing frequency of mixed-effects/multilevel models for corpus data, which allow us to address nested structure in corpus (e.g., the hierarchical structure of corpus sampling protocols) as well as variation between speakers and/or lexical items (see Gries, 2015, 2021, 2024b; Murakami, 2016). However, some additional improvements would still be useful. First, regressions can be followed up in a variety of ways. One very important one of these is referred to as general linear hypothesis (GLH) tests (see Bretz, Hothorn, & Pestfall, 2010). While some scholars now routinely do regression model selection to eliminate insignificant predictors (following Occam's razor) and some do pairwise post hoc comparisons, what is much rarer is the use of GLH tests and/or a priori orthogonal contrasts to determine whether, say, keeping all levels of a categorical predictor distinct is merited or what meaningful groups of levels might 'exist in a predictor'.

Second, for some applications, regression analyses of the above kind can be very fruitfully combined. Gries and Deshors (2014) develop what they call the *MuPDAR(F)* approach (for *Multifactorial Prediction and Deviation Analysis with Regressions(/Random Forests)*). This approach is designed to advance learner corpus research and involves three steps and two regressions (or other predictive modeling techniques):

i. a regression R_1 in which some phenomenon P is studied in native speaker data with a logistic or multinomial regression;
ii. the computation of native-speaker based predictions for learner data;
iii. a regression R_2 which tries to model where the learners did not make the choices the native speakers would have done and why.

Gries and Deshors (2014) show how this approach reveals interesting patterns of how French and Chinese learners' use of *can* and *may* differs from that of native speakers and, more generally, how this approach helps answering one of the most important questions in learner corpus research, namely "given the situation the learner is in, what would a native speaker do?"

Third, a range of other interesting methods can help corpus linguistics tackle other statistical challenges. One example is the approach of Structural Equation Modeling, which is designed to help identify causal effects from correlational effects; see Larsson, Plonsky, and Hancock (2021) for an introduction. Another group of predictive modeling techniques that are becoming more widespread are tree-based methods such as classification and regression trees, conditional inference trees, or random forests, which, with some simplification, involve the construction of flowchart-like tree structures based on successively more fine-grained binary splits of the data; see Strobl, Malley, and Tutz (2009) for a great overview and Heller, Bernaisch, and Gries (2017) or Tomaschek, Hendrix, and Baayen (2018) for recent applications in research on English varieties and on word and segment durations in German respectively.

3.3.2 Exploratory/Hypothesis-Generating Statistics

Apart from modeling/confirmatory approaches discussed so far, there is also a large range of so-called exploratory tools, that is, methods which usually do not test hypotheses and return *p*-values but that detect structure in data that the analyst must then interpret. One of the most widely known methods is of course Biber's multidimensional analysis (MDA); see Biber (1988, 1995) for most comprehensive treatments. In a nutshell, performing an MDA involves (i) annotating a corpus for a large set of relevant linguistic characteristics, (ii) generating a table of normalized frequency counts of each linguistic feature in each part of the corpus, (iii) computing a factor analysis (FA) on this table, which reveals factors/dimensions in the form of co-occurrence patterns of the annotated linguistic features, and (iv) interpreting the co-occurrence patterns in terms of the communicative functions that the co-occurring features perform. MDA has been one of the most influential quantitative methods in

corpus linguistics and has spawned a large number of follow-up studies and replications, many of which used MDA results for the characterization of new registers. In addition, MDA has probably been a main reason for why FA, and its statistical sibling, principal component analysis (PCA), have become popular in corpus linguistic circles long before regression modeling.

Cluster-analytic approaches are also common and they do not require the data to be numeric. Many different kinds of cluster analysis can be distinguished but the most frequent in corpus linguistics is hierarchical agglomerative cluster analysis, which approaches datasets containing *n* items such that it tries to successively amalgamate the *n* items into larger and larger clusters until all items form one cluster; it is then the researcher's task to determine how many clusters there are and what, if anything, they reflect. Other techniques are phylogenetic clustering, which is more flexible than hierarchical clustering in that it does not require all elements to form one cluster at some point, and *k*-means clustering, where the analyst defines the desired/suspected number *k* of clusters and the analysis returns the *n* items grouped into *k* clusters for interpretation; see Moisl (2020) for an overview.

Given their flexibility, cluster analyses can be and have been applied in very many contexts where large and potentially messy datasets were explored for possibly complex correlational structure that would remain invisible to the naked eye. An example involving synchronic corpus and experimental data is Divjak and Gries (2008), who correlate cluster-analytic results of annotated concordance data for nine synonymous Russian verbs with cluster-analytic results from corresponding sorting and gap-filling experiments; an example of using clustering to find temporal structure in diachronic data is Rosemeyer (2015), who shows how priming effects of Spanish auxiliary selection become stronger over the course of 350 years.

Many other exploratory statistical tools are available, too. Examples include (multiple) correspondence analysis (see Glynn, 2010 for an application to the distributional behavior of *bother*), multidimensional scaling (see Sagi, Kaufmann, & Clark, 2011 on how collocates and collocates of *docga*/*dog* and *deo*/*deer* document semantic broadening and narrowing respectively), or network modeling (see Chen, 2025, for a network analysis of degree adverbs in Mandarin corpus data and what it reveals regarding conventionalization from a Construction Grammar perspective).

3.4 Discussion and Concluding Remarks

In spite of having discussed many techniques and desiderata, this chapter could only scratch the surface of quantitative analysis and design in corpus linguistics – most quantitative applications/tools would easily merit an article on their own. However, in addition to the above overview, some cautionary words are also necessary. On top of the mere knowledge of what techniques are available, we also need to make sure that methods are used properly. For

example, it is great if more people use random forests as a more robust alternative to regression modeling, but it is not so great if those are then interpreted on the basis of monofactorial chi-squared tests. We also need firm guidelines on what is important in statistical analysis, what is important to report, and how methods and results should be reported. Other fields have had long and intense discussions about these things – corpus linguistics is still in the beginning stages of this process and we should be inspired by how other disciplines with similar kinds of questions and data have come to grips with these challenges; from my point of view, ecology is most relevant to us, and Wilkinson and the Task Force on Statistical Inference (1999) provide many useful tips.

Given all of the above, it may seem as if corpus linguists are supposed to spend quite some time on learning a large number of sometimes quite complex statistical tests. That perception is entirely accurate. As I have asked elsewhere, why would we as corpus linguists look at something (language) that is based on distributional/frequency-based probabilistic data and just as complex as what psychologists, cognitive scientists, and so on look at but teach in degree programs that often do not contain even a single course on statistical methods? And why would we corpus linguists of course require any kind of medical research to be conducted with the utmost attention to multiple predictors, multiple confounds and controls, their interactions, and variability between subjects but not exercise the exact same care in our own field? If we want to make serious headway in our analyses of corpus data, then, given the immense complexity of our data, we must make serious commitments to learning statistical methodology, and hopefully this chapter succeeded at least in providing an overview of foundational and useful tools that, if adopted, can help us advance our discipline.

References

Aarts, F. G. A. M. 1971. On the distribution of noun-phrase types in English clause-structure. *Lingua*, 26(3), 281–293.

Ben Youssef, C. 2024. mMerge: A corpus driven multiword expressions discovery algorithm. PhD dissertation, University of California, Santa Barbara.

Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. 1995. *Dimensions of register variation*. Cambridge: Cambridge University Press.

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.

Borin, L. 2004. New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. In G. Aston, S. Bernardini, and Dominic Stewart (Eds.), *Corpora and language learners* (pp. 67–87). Amsterdam: John Benjamins.

Bretz, F., Hothorn, T., & Pestfall, P. 2010. *Multiple comparisons using R*. Boca Raton, FL: Chapman & Hall and CRC.

Brook O'Donnell, M. 2011. The adjusted frequency list: A method to produce cluster-sensitive frequency lists. *ICAME Journal*, 35, 135–169.

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.

Chen, A. 2025. From sequentiality to schematization: A two-tier network analysis of covarying collexemes in Mandarin degree adverb constructions. *Corpus Linguistics and Linguistic Theory*, 21(3), 475–515.

Damerau, F. J. 1993. Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management*, 29, 433–447.

Daudaravičius, V., & Marcinkevičienė, R. 2004. Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics*, 9(2), 321–348.

Diessel, H., & Tomasello, M. 2005. Particle placement in early child language: A multifactorial analysis. *Corpus Linguistics and Linguistic Theory*, 1(1), 89–112.

Divjak, D. S., & Gries, St. Th. 2008. Clusters in the mind? Converging evidence from near synonymy in Russian. *The Mental Lexicon*, 3(2), 188–213.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.

Durrant, P., & Schmitt, N. 2009. To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics*, 47(2), 157–177.

Egbert, J., & Biber, D. 2019. Incorporating text dispersion into keyword analyses. *Corpora*, 14(1), 77–104.

Ellis, N. C. 2007. Language acquisition as rational cue-contingency learning. *Applied Linguistics*, 27(1), 1–24.

Ellis, N. C., & Ferreira-Junior, F. 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, 7, 187–220.

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

Glynn, D. 2010. Testing the hypothesis: Objectivity and verification in usage-based cognitive semantics. In D. Glynn & K. Fischer (Eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches* (pp. 239–629). Berlin: De Gruyter Mouton.

Gries, St. Th. 2000. Multifactorial analysis in corpus linguistics: The case of particle placement. PhD dissertation, University of Hamburg.

Gries, St. Th. 2003. *Multifactorial analysis in corpus linguistics: A study of particle placement*. London: Continuum Press.

Gries, St. Th. 2005. Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory*, 1(2), 277–294.

Gries, St. Th. 2006. Exploring variability within and between corpora: Some methodological considerations. *Corpora*, 1(2), 109–151.

Gries, St. Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), 403–437.

Gries, St. Th. 2012. Frequencies, probabilities, association measures in usage-/exemplar-based linguistics: Some necessary clarifications. *Studies in Language*, 36(3), 477–510.

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

Gries, St. Th. 2015. The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora*, 10(1), 95–125.

Gries, St. Th. 2021. *Statistics for linguistics with R*. 3rd rev. & ext. ed. Boston: De Gruyter.

Gries, St. Th. 2022a. Multi-word units (and tokenization more generally): A multi-dimensional and largely information-theoretic approach. *Lexis*, 19. https://doi.org/10.4000/lexis.6231.

Gries, St. Th. 2022b. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies*, 5(2), 171–205.

Gries, St. Th. 2024a. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam: John Benjamins.

Gries, St. Th. 2024b. Against level-3-only analyses in corpus linguistics. *ICAME Journal*, 48(1), 1–25.

Gries, St. Th., & Deshors, S. C. 2014. Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora*, 9(1), 109–136.

Gries, St. Th., & Hilpert, M. 2010. Modeling diachronic change in the third person singular: A multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics*, 14(3), 293–320.

Gries, St. Th., & Hilpert, M. 2012. Variability-based neighbor clustering: A bottom-up approach to periodization in historical linguistics. In T. Nevalainen & E. Closs Traugott (Eds.), *The Oxford handbook of the history of English* (pp. 134–144). Oxford: Oxford University Press.

Gries, St. Th, & Mukherjee, J. 2010. Lexical gravity across varieties of English: An ICE-based study of *n*-grams in Asian Englishes. *International Journal of Corpus Linguistics*, 15(4), 520–548.

Gries, St. Th., & Stefanowitsch, A. 2004. Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, 9(1), 97–129.

Han, W, Arppe, A., & Newman, J. 2017. Topic marking in a Shanghainese corpus: From observation to prediction. *Corpus Linguistics and Linguistic Theory*, 13(2), 291–319.

Heller, B., Bernaisch, T., & Gries, St. Th. 2017. Empirical perspectives on two potential epicenters: The genitive alternation in Asian Englishes. *ICAME Journal*, 41, 111–144.

Hilpert, M. & Gries, St. Th. 2009. Assessing frequency changes in multi-stage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 34(4), 385–401.

Hinrichs, L., & Szmrecsanyi, B. 2007. Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics*, 11(3), 437–474.

Ikehara, S., Shirai, S., & Uchino, H. 1996. A statistical method for extracting uninterrupted and interrupted collocations from very large corpora. *Proceedings of the 16th Conference on Computational linguistics*, 1, 574–579.

Kendall, T., Bresnan, J., & van Herk, G. 2011. The dative alternation in African American English researching syntactic variation and change across sociolinguistic datasets. *Corpus Linguistics and Linguistic Theory*, 7(2), 229–244.

Kita, K, Kato, Y., Omoto, T., & Yano, Y. 1994. Automatically extracting collocations from corpora for language learning. *Journal of Natural Language Processing*, 1(1), 21–33.

Larsson, T, Plonsky, L., & Hancock, G. 2021. On the benefits of structural equation modeling for corpus linguists. *Corpus Linguistics and Linguistic Theory*, 17(3), 683–714.

Leech, G., & Fallon, R. 1992. Computer corpora: What do they tell us about culture? *ICAME Journal*, 16, 29–50.

Leech, G. N., Francis, B., & Xu, X. 1994. The use of computer corpora in the textual demonstrability of gradience in linguistic categories. In C. Fuchs & B. Victorri (Eds.), *Continuity in linguistic semantics* (pp. 57–76). Amsterdam: John Benjamins.

Mair, C., Hundt, M., Leech, G., & Smith, N. 2002. Short term diachronic shifts in part-of-speech frequencies: a comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics*, 7(2), 245–264.

Mason, O. 1997. The weight of words: An investigation of lexical gravity. In B. Tomaszczyk & P. J. Melia (Eds.), *Proceedings of PALC'97* (pp. 361–375). Lodz: Lodz University Press.

Mason, O. 1999. Parameters of collocation: The word in the centre of gravity. In J. M. Kirk (Ed.), *Corpora galore: Analyses and techniques in describing English* (pp. 267–280). Amsterdam: Rodopi.

McCauley, S. M., & Christiansen, M. H. 2019. Modeling children's early linguistic productivity through the automatic discovery and use of lexically-based frames. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual conference of the Cognitive Science Society* (pp. 782–788). Austin, TX: Cognitive Science Society.

Michelbacher, L., Evert, S., & Schütze, H. 2011. Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory*, 5(1), 79–103.

Moisl, H. 2020. Cluster analysis. In M. Paquot & St.Th. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 401–434). Berlin: Springer.

Murakami, A. 2016. Modeling systematicity and individuality in nonlinear second language development: The case of English grammatical morphemes. *Language Learning*, 66(4), 834–871.

Nagao, M., & Mori, S. 1994. A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese. *Proceedings of the 15th Conference on Computational Linguistics*, Vol. 1 (pp. 611–615). https://aclanthology.org/C94-1101.pdf.

Onnis, L., & Thiessen, E. 2012. Language-induced biases on human sequential learning. Paper presented at the 34th Annual Meeting of the Cognitive Science Society, Sapporo, Japan.

Paquot, M., & Bestgen, Y. 2009. Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In A. Jucker, D. Schreier, & M. Hundt (Eds.), *Corpora: Pragmatics and discourse* (pp. 247–269). Amsterdam: Rodopi.

Pecina, P. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1–2), 137–158.

R Development Core Team. 2024. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. www.R-project.org.

Rosemeyer, M. 2015. How usage rescues the system: Persistence as conservation. In A. Adli, M. García, & G. Kaufmann (Eds.), *Variation in language: System- and usage-based approaches* (pp. 289–316). Boston: De Gruyter.

Sagi, E., Kaufmann, S., & Clark, B. 2011. Tracing semantic change with latent semantic analysis. In J. Robynson & K. Allan (Eds.), *Current methods in historical semantics* (pp. 161–183). Berlin: Mouton de Gruyter.

Schmitt, N. & Redwood, S. 2011. Learner knowledge of phrasal verbs: A corpus-informed study. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora: In honor of Sylviane Granger* (pp. 173–207). Amsterdam: John Benjamins.

Strobl, C., Malley, J., & Tutz, G. 2009. An introduction to recursive partitioning: rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods*, 14(4), 323–348.

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.

Tomaschek, F., Hendrix, P., & Baayen, R.H. 2018. Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics*, 71, 249–267.

Wilcox, R. R. 2017. *Understanding and applying basic statistical methods using R*. Hoboken, NJ: John Wiley & Sons.

Wilkinson, L. & the Task Force on Statistical Inference. 1999. Statistical methods in psychology journals: Guidelines and expectations. *American Psychologist*, 54(8), 594–604.

Williams, J., & Reagan, A. 2017. Python partitioner. Version 0.1.2, https://github.com/jakerylandwilliams/partitioner.