Chapter

Statistics in corpus linguistics

Stefan Th. Gries

Abstract

The primary goal of this paper is to provide an overview of the use of statistical methods in corpus linguistics in the hope that readers will be able to, after having read this chapter, understand original corpus-linguistic research studies and their methodological choices and can begin to consider choices for their own applications; a secondary goal is to augment the overview with some necessarily subjective and critical discussion and suggestions for ways current research practices might be improved. In Section 2, I discuss a range of specifically corpus-linguistic statistics having to do with frequency, dispersion, and association and I show at least briefly how these notions can be measured/operationalized statistically and some of the concerns and pitfalls users need to be aware of. In Section 3, I

offer a brief survey of a variety of statistical methods ranging from descriptive statistics, statistical modeling/machine learning, and exploratory tools and briefly discuss a few studies that highlight some of these methods' strengths.

1. Introduction

Much corpus-linguistic work is ultimately based on a combination of (i) 1+ several different corpus retrieval operations and (ii) 1+ statistical operations. The retrieval operations of (i) can be distinguished in terms of how much context they involve: For instance, if the goal is

- a **frequency list** of a corpus (e.g., to know how frequent each word is in the corpus), which requires retrieving all words from a corpus but does not require contexts of those words – their acontextual frequency alone is sufficient; this, as many things below, is a bit of a simplification since it ignores, for example, the often tricky problem of how to deal with multi-word units such as *because of*: if one wanted to be able to count instances of that expression as opposed to just counting instances of *because* and *of*, some contextually sensitive parsing of the corpus into words (a process called tokenization) would be required;
- **dispersion statistics** for a corpus (e.g., to know how evenly distributed a word is in the corpus), which requires retrieving all words from a corpus as well as which corpus part/file they occur in how often (and often the sizes of the corpus parts) – but the words' contexts within sentences or utterances are not required;
- a **collocation/collostructional study** of co-occurring elements (e.g., to know how much a word *w* 'likes to co-occur with' another word or a construction), which requires retrieving all instances of *w* from a corpus and the relevant other words/constructions in, typically, a small context defined by a window of words or a syntactically-defined slot;
- a **concordance** (e.g., to know exactly how, say, a word *w* is used), which requires retrieving all instances of *w* in their complete context.

Since each of these routes ultimately leads to frequencies – the frequencies of a word in a corpus or its parts, the frequencies of collocates around a word, the frequencies of any kinds of contextual features around a word – corpus linguists more often than not have to deal with statistical methods to address the questions they are interested in. Such statistical methods can be heuristically considered as coming in two different kinds: They might require

- specifically **corpus-linguistic statistics**: such as different kinds of frequencies, dispersion statistics, and association measures (to quantify the degree of attraction or repulsion of an element to (i) either some other linguistic element in the context (collocational/collostructional studies) or to (ii) one of two or more corpora they are attested in (keyness studies quantifying how characteristic a word is for a corpus);
- **general statistical methods** that can be applied to any kind of data: ecological data, psychological data, …, and corpus-linguistic data.

This overview surveys and exemplifies both kinds of statistical methods that are often found in corpus studies. Section 2 deals with specifically corpus-linguistic statistics: frequencies, dispersion, and association/keyness; it mentions some of the most important considerations going into the choice and use of such measures and exemplifies some of them in the R programming language and environment (see https://cran.r-project.org); this is to give readers an idea of how relatively simple it is to compute such measures without having to depend on custom-made limiting software applications. Section 3 then turns to general statistical methods and discusses general descriptive statistics, the two main kinds of statistical modeling techniques used in state-of-the-art studies, and then some exploratory methods for situations in which no specific hypotheses are tested; given the complexity of many of these methods, in this section, I can only provide general characteristics of these methods, but I will provide plenty of references for future reference. Section 4 concludes.

2. Corpus-linguistic statistical measures

2.1 Frequency

2.1.1 Overview

The most basic corpus-linguistic statistic is frequency of occurrence, which usually comes in two kinds: **Token frequencies** state how often a certain word, lemma, construction, morpheme, etc. type is attested in a corpus. For instance, consider the following lines of code in the R programming language (R Core Team, 2021) that creates a small 'schematic corpus' (called words) consisting of five parts *p1* to *p5*, in which each letter represents a different word:

```
corpus.parts <- paste0("p", rep(1:5, c(9,10,10,10,11)))
words <-
c("b","a","m","n","i","b","e","u","p","b","a","s","a","t","b","e","w","q","n"
,"b","c","a","g","a","b","e","s","t","a","b","a","g","h","a","b","e","a","a",
"t","b","a","h","a","a","b","e","a","x","a","t")
tapply(words, corpus.parts, noquote)
## $p1
## [1] b a m n i b e u p
##
## $p2
##  [1] b a s a t b e w q n
##
## $p3
##  [1] b c a g a b e s t a
##
## $p4
##  [1] b a g h a b e a a t
##
## $p5
##  [1] b a h a a b e a x a t
```

In this corpus, the token frequency of the word *a* is 15, which we can quickly determine from either a full frequency table of the corpus as created with table or from just counting the number of *a*s in the corpus directly:

```
table(words)
## words
##  a  b  c  e  g  h  i  m  n  p  q  s  t  u  w  x
## 15 10  1  5  2  2  1  1  2  1  1  2  4  1  1  1
sum(words=="a")
## [1] 15
```

The second important kind of frequency is **type frequency**, which states how many *different* tokens (i.e. types) are attested in a corpus or in a slot around a word / of a construction; the above toy corpus has a word type frequency of 16:

```
length(unique(words))
## [1] 16
```

Both token and type frequencies can be reported as **absolute frequencies**, which are the raw observed numbers as given above, but often they are relativized/normalized to something like 100,000 or 1 million words; such **relative frequencies** permit comparisons of token frequencies from differently large corpora. For example, the word *c* occurs once in the above corpus of 50 words, meaning its frequency per million words (pmw) can be computed as follows:

$$\frac{observedtokenfrequency}{corpussizeintokens} \times 1000000 = \frac{1}{50} \times 1000000 = 20000$$

It is worth pointing out, however, that, while this kind of frequency is often reported, it is not without risks: For many linguistic elements under consideration, a normalization by the number of words (i.e. putting the corpus size in words into the denominator like we did here) is not obviously the right choice. For instance, if the focus of a study is on 'something morphemic', relative frequencies based on words will be less than ideal and putting something more closely approximating the total token frequency of morphemes in the corpus will fare better; same if the focus of the study is on 'something syntactic/constructional'. Thus, the notion of 'corpus size' is one that needs to be operationalized carefully.

A fairly recently suggested improvement to such frequencies is the **Zipf scale** (van Heuven et al., 2014), another measure aiming at making frequencies from different corpora more comparable. This measure comes in two versions. The 'basic' one is simply computed like this:

$$Zipfscale = 3 + log_{10} obs.freq._{pmw}$$

For the word *c* above, this would mean the Zipfscale-value would be 7.30103:

```
3+log10(20000)
## [1] 7.30103
```

The 'more advanced' version of the Zipfscale also takes into consideration a number of word types that were not actually observed in a corpus but might have been. It is computed as follows:

$$Zipfscale = 3 + log_{10}\frac{obs.freq.+1}{\left(corpussize_{tokens} + corpussize_{types}\right) \div 1000000}$$

For the 'word' *c* above, this would mean the Zipfscale-value would be 7.481486:

```
3+log10(2/((50+16)/1000000))
## [1] 7.481486
```

In addition to these frequencies of occurrence, we also find **frequencies of co-occurrence**, which may also be expressed in two ways. One might report an absolute frequency such as 'in the above toy corpus, the collocation *a g* occurs two times':

```
length(intersect(
  which(words=="a"),
  which(words=="g")-1))
## [1] 2
```

Alternatively, one might report a relative frequency or conditional probability: Since there are 15 occurrences of *a* and 2 occurrences of *g* in the corpus, the collocation *a g*'s relative frequency (relative to *a*) could be expressed as $^{2}/_{15} = 0.1333333$ while *a g*'s relative frequency (relative to *g*) could be expressed as $^{2}/_{2} = 1$.

2.1.2 Applications/discussion

Frequencies of occurrence and/or co-occurrence are relevant in different fields and for different reasons. In more applied settings, frequencies inform pedagogical applications (e.g., which words/patterns to teach (first/early)). In cognitive-linguistic and psycholinguistic settings, the frequency of, say, a word has been argued to be a useful proxy of its 'commonness in a language/dialect' and of its degree of cognitive entrenchment, which in turn

is correlated with ease and speed of access and comprehension as measured by, for instance, reaction times in lexical decision/ recognition tasks. Correspondingly, psycholinguistic models accommodate frequency effects in various ways, e.g., in strengths of connections between nodes representing lexical items or grammatical structures, in (higher) resting levels of activation, etc. In many diachronic linguistics applications, frequencies are important because, e.g., elements with a high token frequency resist diachronic regularization patterns more than rarer elements and expressions with high type frequencies in one or more of their slots – i.e., here we are including also co-occurrence frequency of, say, a construction and something in one of the construction's slots – are more prone to grammaticalize and, thus, take on more general meanings. In first language acquisition, high token frequency of exposure is correlated with age and ease of acquisition and high type frequencies in slots of constructions are correlated with children's ability to generalize and form linguistic categories. In many different linguistic areas, high token frequencies are also probabilistically correlated with, though not fully determinative of, the status of an element as the prototype of its category, etc.[1]

However, in spite of all these well-documented correlations, two crucial problems remain: First, correlation does not prove causation and since frequency is not just correlated with response variables such as reaction times etc. but also correlated with other predictors of the same responses (e.g., rated familiarity, concreteness, age-of-acquisition, word length, …), it is a non-trivial task to determine what role frequency really plays. Studies such as McDonald and

[1] As an aside: it would of course be possible to actually consider all such frequencies of occurrence as frequency of co-occurrence because (i) frequencies of occurrence of, for example, a word in a construction by definition correspond to a co-occurrence frequency of that word and that construction and (ii) while frequency counts are usually based on the formal aspects of a linguistic sign (often words), each use of a word of course (co)occurs with a certain (semantic, discoursal, …) sense/function, but I will not pursue this more 'philosophical' argument here.

Shillcock (2001) or Baayen (2010) have demonstrated that frequency as a repetition counter might in fact be much less causally related with the above response variables and that other explanatory variables/predictors are in fact more useful; this will probably be one of the most important areas of research in the short-to medium-term future.

Second, even though frequencies are probably the most widely-used corpus statistic, they are often highly problematic, especially if they are used to represent the 'widespreadedness' or 'commonness' of a word in a register or a corpus, which is in fact their most frequent application. That is because any such frequency statistic is essentially a mean without a dispersion statistic indicating how well the frequency/mean represents the distribution it tries to represent – for this, we need dispersion measures, which will be discussed now.

2.2 Dispersion

2.2.1 Overview

To highlight the importance of dispersion, consider the following question: How would you rank the following words in terms of commonness in spoken data (as represented by the spoken part of the British National Corpus (BNC)): *council*, *nothing*, *try*, *whether*? I am assuming everyone would consider *council* the outlier and think that *council* is less common/widespread than the rest: One would expect *council* to be less frequent, more clumpily distributed in a corpus (especially one of spoken language), to be acquired later by children and/or learners, and to be less polyfunctional than the others, … But, as we can see in Table 1, all four words have extremely similar frequencies in the spoken component of the BNC:

Table 1: Frequencies of four words in the spoken part of the BNC

Word	*council*	*nothing*	*Try*	*whether*
Token frequency	4387	4159	4199	4490

However, this does not disprove any analyst's likely intuition that *council* would be less common – it might just as well suggest that

frequency is a potentially problematic measure (especially if not contextualized with a dispersion measure). Once we consider these words' dispersions by, for instance, determining the absolute number of different files the words are attested in at least once – a crude dispersion measure called *range* – an analyst's likely intuition is in fact confirmed:

Table 2: Ranges of four words in the spoken part of the BNC

Word	*council*	*nothing*	*Try*	*whether*
Dispersion: range	292	652	664	671

Now a reader might of course claim that, for whatever (sampling) reason, *council* is an exception – and maybe it is. However, one can avoid having to resort to such a rhetorical slight of hand simply by operationalizing 'commonness' better, namely with some measure of dispersion, which here 'gives the intuitively right answer'. Crucially, this kind of situation is more common than one might think: In the 1m-words Brown corpus (consisting of 500 samples of written American English), the words *enormous* and *staining* have the exact same absolute frequency of occurrence (37), but the instances of *enormous* are spread out over 36 different corpus parts whereas all instances of *staining* are from only 1 of the 500 corpus parts – given this distributional difference, claiming that both words are equally common (because their frequency is the same) makes no sense at all.

How can dispersion be measured? The most primitive measure is the measure of *range* just discussed, which, however, is usually expressed not as an absolute number of corpus parts but as the percentage of the corpus parts that a word of interest is attested in one or more times. In the case of the above toy corpus, we can compute this measure quickly from a **term-document matrix**, a table that has all word types in the rows, all corpus parts in the columns, and the occurrences-per-file in the main cells:

```
(tdm <- table(    # make tdm a table with
  words,          # the words in the rows
  corpus.parts)) # the corpus parts in the columns
##      corpus.parts
## words p1 p2 p3 p4 p5
##     a  1  2  3  4  5
##     b  2  2  2  2  2
##     c  0  0  1  0  0
##     e  1  1  1  1  1
##     g  0  0  1  1  0
##     h  0  0  0  1  1
##     i  1  0  0  0  0
##     m  1  0  0  0  0
##     n  1  1  0  0  0
##     p  1  0  0  0  0
##     q  0  1  0  0  0
##     s  0  1  1  0  0
##     t  0  1  1  1  1
##     u  1  0  0  0  0
##     w  0  1  0  0  0
##     x  0  0  0  0  1
```

We just need to count for each word/row how many of the frequencies per corpus part/column are greater than 0 and divide that number by the number of corpus parts (5), and then we can sort these percentages to see that, for instance, the range of the word *a* is 1 (it occurs in every part) and the range of the word *c* is 0.2 (it occurs in only 1 of 5 parts):

```
sort(ranges <- apply( # apply to
  tdm,                # the term-document matrix,
  1,                  # namely, each row
  # an anonnymous function that checks how many frequencies are >0
  \(af) sum(af>0)) / 5)
##   c   i   m   p   q   u   w   x   g   h   n   s   t   a   b   e
## 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.4 0.4 0.4 0.4 0.8 1.0 1.0 1.0
```

The probably most widely-used measure of dispersion is Juilland's *D*, but recent comparative studies by Biber et al. (2016) and Burch et al. (2017) have shown that Juilland's *D* is actually somewhat problematic, which is why I do not explain its computation here (see Gries 2020 for the formula), and that Gries's *DP* is superior at least in their applications (it also outperforms Juilland's *D* in Gries

2010). With a term-document matrix like tdm, *DP* is extremely easy to compute:

1. one converts the frequencies with which a word occurs in each corpus into row proportions (i.e. proportions of the overall frequency of the word);
2. one converts the corpus part sizes into proportions (of the overall corpus size);
3. one computes the pairwise differences of these proportions, takes their absolute values, sums them up, and divides by 2.

A potential 4th step could be to normalize *DP* to DP_{norm} by dividing *DP* by 1-the smallest corpus part size (to make sure the values exhaust the interval of [0, 1]). Here, we do this for word *a*:

```
(step.1 <- tdm["a",]/sum(tdm["a",]))
##         p1         p2         p3         p4         p5
## 0.06666667 0.13333333 0.20000000 0.26666667 0.33333333
# meaning, 6.6667% of a is in the 1st corpus part
(step.2 <- colSums(tdm)/sum(tdm))
##   p1   p2   p3   p4   p5
## 0.18 0.20 0.20 0.20 0.22
# meaning, the first corpus part is 18% of the corpus
(DP <- sum(abs(step.1-step.2))/2) # computing DP from pairwise differences
## [1] 0.18
(DP.norm <- DP/(1-min(step.2))) # normalizing it to [0, 1]
## [1] 0.2195122
```

Note that *DP*'s and DP_{norm}'s orientation is such that high values mean clumpy/uneven distribution whereas low values mean even distribution – if the opposite orientation is desired, one can just use 1-$DP_{(norm)}$. Here are the *DP*-values for the above 4 words from the spoken part of the BNC, and again we see that *council* is much more clumpily distributed than the other three words:

Table 3: *DP*-values of four words in the spoken part of the BNC

Word	*council*	*nothing*	*Try*	*whether*
DP	0.7178632	0.2802748	0.2802816	0.3155424

Thus, given (i) how differently dispersed even words with the same frequency can be, (ii) how there is mounting evidence that words' commonness might be better approximated by dispersion than by frequency, and (iii) the relative ease with which dispersion can now be computed, there is really no good excuse anymore not to use it.

2.2.2 Applications/discussion

Given the above, it will not come as a surprise that I would basically argue that dispersion is relevant in most cases in which researchers have so far restricted themselves to frequency. This is for two main reasons: First, frequency and dispersion simply answer different questions and it seems to me that, while frequency is easy to compute and seems straightforward to integrate into our explanations/theories, for many applications the question that dispersion answers is actually more pertinent. Frequency answers the question "how often does *x* happen?" whereas dispersion answers the question "in how many contexts/situations will you encounter *x*?". This not only establishes a clear connection to all sorts of recency effects in language, memory, and processing (see e.g. Ambridge et al., 2006, p.175), but we now have a growing body of evidence that shows that dispersion's predictive power is higher than that of frequency. The above-mentioned studies of Baayen (2010) and Gries (2010) showed that dispersion metrics have a higher degree of predictive power than frequency when it comes to lexical decision times; Adelman et al. (2006) offer similar results (but seem unaware that what they are testing is dispersion); Ellis and colleagues have shown that range has a significant predictive power when it comes to construction uptake beyond raw frequency (Ellis & Simpson-Vlach, 2005; Ellis et al., 2007).

Second, frequencies of (co-)occurrence underlie pretty much all corpus statistics, but since frequencies fail to reflect matters of dispersion, using frequencies without accompanying dispersion measures can lead to very misleading results. For instance, as early as 2003, Stefanowitsch & Gries showed that an association measure based only on corpus frequencies indicates that the verbs *fold* and *process* are strongly attracted to the imperative construction in the

British Component of the International Corpus of English (ICE-GB) – however, a closer look revealed this to be a bit of an artifact because these results are due to only a single file each in the whole corpus (one book on Origami and one cookbook respectively). In other words, any analysis based on frequency alone runs the risk of reporting results that are completely skewed by one total outlier.

Bottom line, researcher should *always* inspect dispersion statistics for *any* statistics computed from corpus frequencies: Not only might those have a higher predictive power than anything based on frequencies anyway, but this would also insure the researchers against jumping to conclusions based on outliers. However, including dispersion statistics can be done in two ways: (i) one can merge one's frequency-based statistic(s) with dispersion information into a single (nicely sortable) vector of values or (ii) one can consider both dimensions of information – frequency and dispersion – at the same time yet separately. In the following excursus, I will argue that the former is not uncommon, but ultimately misguided.

2.2.3 Excursus: frequency and dispersion

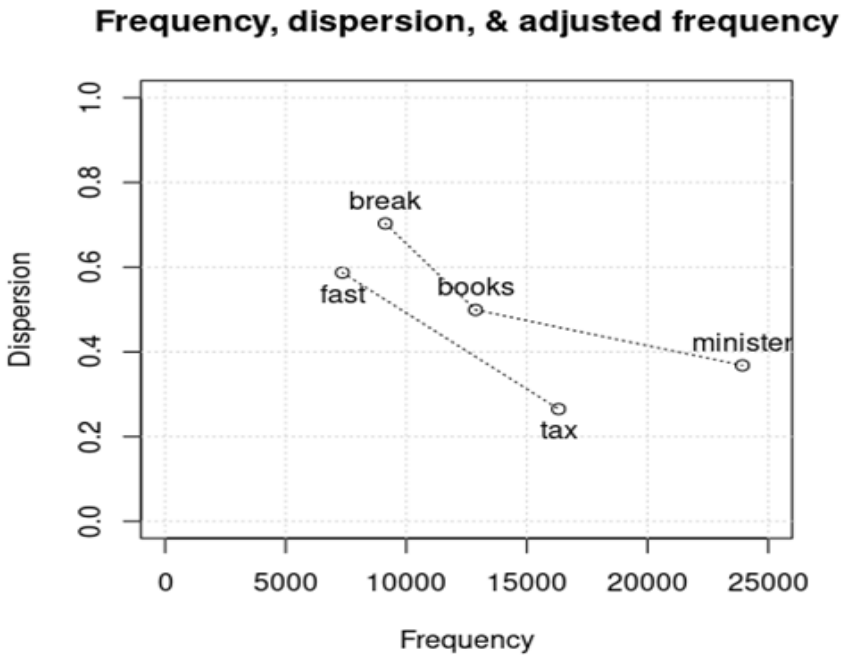
Sometimes, one finds applications where researchers are aware of the relevance of dispersion information for a particular application and compute it in addition to, for instance, observed frequencies. In lexicography, for instance, dispersion is sometimes used to adjust frequency information such that

- if a word *w* is distributed very unevenly/clumpily in a corpus (like *staining* in the Brown corpus), its frequency gets adjusted downwards considerably (to avoid reporting a perhaps fairly high observed frequency of a word that is actually only attested in a very small part/section of the corpus);
- if a word *w* is distributed more evenly in a corpus (like *enormous* in the Brown corpus), its frequency gets adjusted downwards much less or not at all (because the even spread of the word throughout the corpus lends more credibility to the overall observed frequency).

The adjustment of an adjusted frequency might simply consist of multiplying the observed token frequency with a dispersion measure whose orientation is such that low and high values reflect clumpy and even distributions respectively (such as 1-$DP_{(norm)}$). For the four words in the spoken part of the BNC discussed above that means that the frequency of *council* would be adjusted downwards quite a bit (because *council* is so clumpily distributed) whereas the frequency of the other three words would be adjusted downwards much less (because they are so much more evenly distributed). In other words, researchers might compute the observed the frequencies and the dispersions of words, but then compute the corresponding adjusted frequencies and only report those. However, that conflation of two dimensions (frequency and dispersion) for each word into one dimension (an adjusted frequency) is not a good idea because of the inevitable and massive information loss it incurs. Consider these three words and their extremely similar adjusted frequencies in the complete BNC: *break* (adj. freq.: 6419), *books* (adj. freq.: 6420), and *minister* (adj. freq.: 6415), or these two words in the same corpus: *fast* (adj. freq.: 4317) and *tax* (adj. freq.: 4316). But these very similar adjusted frequencies are from very different actual frequencies and dispersion values, as shown in Table 4 and visually represented in Figure 1 (where words are plotted at coordinates of their frequency and dispersion and words connected by lines have the same adjusted frequency):

Table 4: Adjusted & Observed Frequencies of 5 Words in the BNC and Their Dispersions

Word	Observed frequency	Dispersion	Adjusted frequency
break	9128	0.703	6419
books	12872	0.499	6420
minister	23935	0.368	6415
fast	7349	0.587	4317
tax	16313	0.265	4316



with completely different distributional characteristics seem virtually identical; it's hard to imagine an application where this massive loss of information would be useful, which is why keeping frequency and dispersion separate (as in the left three columns of Table 4 and Figure 1) is by default a better way to go.

2.3 Association and keyness

2.3.1 Overview

Maybe the most central assumptions underlying much of corpus linguistics is the **distributional hypothesis**, here in the form provided by Harris (1970:785f.):

> [i]f we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. […], difference of meaning correlates with difference of distribution.

In other words, distributional similarity reflects functional

similarity, where *functional* is broadly construed as encompassing one or more of semantic, discourse-functional, information-structural, and other kinds of similarity. That implies that words that are semantically similar tend to occur in similar lexical and grammatical contexts. For instance, the **collocates** (i.e. the words you find 'around') of the noun *cat* will be more similar to the collocates of the noun *dog* than to the collocates of the adjective *ethereal*. This distributional hypothesis has been used particularly much in studies of near synonymy of both lexical and syntactic constructions, i.e. for sets of words or syntactic constructions with extremely similar meanings/functions. While most native speakers of English might not be able to explain to a learner of English what the precise semantic differences are between *brisk*, *fast*, *quick*, *rapid*, *speedy*, and *swift* or between *deadly*, *fatal*, *lethal*, *mortal*, they usually experience no difficulties whatsoever deciding which to use in natural conversation and they do so fairly consistently, and one corpus-linguistic way of trying to tease apart such near synonyms would involve looking at the words that co-occur with them within a certain user-defined window (e.g. 4L-4R, meaning 'from four words to the left to four words on the right') or within a certain user-defined grammatical slot (e.g. the nouns that follow, and are modified by, these adjectives). This logic extends to the co-occurrence of words and constructions as well: In applications of **collostructional analysis**, many constructions have been shown to have strong preferences for certain (classes of) lexical items to fill their slots; for example, the ditransitive construction [$_{VP}$ … [$_{NPrecipient}$ …] [$_{NPpatient}$ …]] has been associated with the meaning of 'transfer' (of the patient from an agent to the recipient) and the verb slot of this construction is indeed highly associated with verbs of transfer (in particular *give*, see Stefanowitsch & Gries, 2003; Gries & Stefanowitsch, 2004).

How are such associations reported? The simplest way would be frequencies again:

- absolute co-occurrence frequencies: how often does one find the collocation *hermetically sealed* in a corpus or how often does one find *regard* in the *as*-predicative ([$_{VP}$ V

$[_{\text{NPdirectobject}}$ …] *as* [~XP …]])?

- relative frequencies / conditional probabilities: how much of the uses of *hermetically* (in percent) is followed by *sealed* in a corpus or how much of the uses of the *as*-predicative (in percent) contains the verb *regard* in its verb slot?

However, most of the time such associations are quantified using dedicated **association measures** (AMs), most of which are based on 2×2 tables such as Table 5, which cross-tabulates the frequencies of the two elements in whose association one is interested in:

Table 5: A Schematic Co-Occurrence Table Underlying Nearly All Widely-Used Association Measures

	Element 2	**Not element 2**	**Sum**
Element 1	*a*	*b*	*a+b*
Not element 1	*c*	*d*	*c+d*
Sum	*a+c*	*b+d*	*N*

In such a table,

- the cell *a* would represent the absolute token frequency of co-occurrence of elements 1 and 2;
- the cell *b* would represent the absolute token frequency of element 1 without element 2;
- the cell *c* would represent the absolute token frequency of element 2 without element 1;
- the cell *d* would represent the absolute token frequency of co-occurrence of not element with not element 2. Note that as per the comment on what goes into the denominator of relative frequencies in Section 2.1 above (words, morphemes, syntactically-defined slots?), determining the right unit for the cells *d* and, thus, *N* is not always uncontroversial – what is, or how do we count, 'not element 1' and 'not element 2'? The answer to this question must be tailored to the nature of the question as well as possible. Thus,

- for lexical co-occurrence/collocations, *N* is usually the corpus size in word tokens;
- for a collostructional study of an argument structure construction, where such constructions usually involve a lexical verb as their core element in one of their slots, *N* has often been approximated with the token frequency of lexical verbs.

This topic has been discussed especially in the context of collostructional studies.[2] However, regardless of the nature of the elements, uses of AMs nearly always follow a certain four-step template:

1. one retrieves (ideally) all instances of a first element of interest, say a construction *C*;
2. for each type of the second element of interest (e.g., a verb in a slot of *C*), one computes an AM that is (usually) based on the relevant 2×2 tables of the above kind;
3. one sorts the second elements of interest according to that AM;
4. one analyzes the top *x* elements of interest in terms of their structural, semantic, or other functional characteristics.

As an example of the kind of findings one might obtain for the above speed adjectives, consider some of the most attracted noun collocates after each speed adjective in the BNC (searched for without tags and loosely grouped together by word family and semantics):

- *fast*: *bowler/bowlers/bowling*, *food*, *lane*, *track*, *car/cars*, *reactor/reactors*, *breeder*, *pace*, *buck*, …;
- *quick*: *glance*, *look*, *succession*, *fix*, *throw-in*, *thinking*, *smile*, *response*, *word*, *reference*, …;

[2] See Bybee (2010, p.98) for critical discussion of the *d*-cell in collostructional studies and Gries (2012, p.487f.) for empirical results showing that Bybee exaggerated the size of the problem, followed by Schmid and Küchenhoff (2013) and yet another rebuttal by Gries (2015).

- *rapid*: *growth*, *expansion*, *rise*, *decline*, *succession*, *development*, *change/changes*, *progress*, *deployment*, *rate*;
- *swift*: *movement*, *glance*, *action*, *investment*, *tuttle*, *return*, *kiss*, *recovery*, *rise*, *response*.

While there is some overlap (and while this is only a small selection of the most-attracted collocates), some semantic observations emerge, e.g. the connection of *fast* with bowling and cars, the connection of *quick* to mental/communication acts, the connection of *rapid* to development in abstract domains, etc.

But how does one compute AMs that give rise to such findings? With some simplification, two kinds of AMs can be computed for such tables: **bidirectional measures** that quantify the *mutual* attraction between the two elements in question or **unidirectional measures** that quantify how much one element (1 or 2) attracts/repels the other (2 or 1) but *not* vice versa.

2.3.2 The standard (still): bidirectional measures

For example, for a collocational study – e.g., what are the nouns that follow the word *fast*? – this would involve creating for each nominal collocate of *fast* (e.g., *car*) a version of the following table (using made-up frequencies for exemplification only):

Table 6: A Concrete Co-Occurrence Table for "Fast Car"

	car	**Not *car***	**Sum**
fast	100	900	1000
Not *fast*	400	98600	99000
Sum	500	99500	100000

Such a table could be entered into R as follows:

```
(fast.car <- matrix(c(100, 400, 900, 98600), ncol=2, dimnames=list(
  ADJ=c("fast", "others"), NOUN=c("car", "others"))))
##         NOUN
## ADJ      car others
##   fast   100    900
##   others 400  98600
```

The most commonly-used measures one needs to know when reading about collocational studies seem to be the following:

- pointwise Mutual Information (*MI*), a value in the interval [-", +"] where positive numbers mean 'mutual attraction' and negative numbers mean 'mutual repulsion';
- the loglikelihood ratio (G^2), a value in the interval [0, +"] whose size is correlated with the absolute value of *MI* and indicates how much the distribution differs from chance (and one would need to look at, say, the sign of the *MI*-score to see whether the two words attract or repel each other);
- the odds ratio (*OR*), a value in the interval [0, +"] where, if the *OR* is computed as below, values >1 mean the elements in the first row and first column attract each other and where values <1 mean the elements in the first row and first column repel each other. Sometimes, this value might be logged, in which case attraction/repulsion of the elements in the first row and first column is represented by positive/negative values respectively.

For out example of Table 6, these three measures could be computed in R as follows:

```
# MI for fast car
log2(100 / (1000*500/100000))
## [1] 4.321928
# G2 for fast car
glm(fast.car ~ c("car", "not car"), family=binomial)$null.deviance
## [1] 438.1371
# OR for fast car
(100/900) / (400/98600)
## [1] 27.38889
```

All of these point to a strong mutual association between *fast* and *car*: For example and using the odds ratio, if *fast* is there, then the odds of *car* are 1 to 9 (100 vs. 900), but if *fast* is not there, the odds of *car* are only 1 to 246.5 (400 vs. 98600), indicating how

much less surprising the presence of *car* is when one has already seen *fast*.

For a collostructional study – e.g., what are the verbs that like to occur in the verb slot of the ditransitive construction? – this would involve creating for each verb occurring in the construction (e.g., *give*) the following table:

Table 7: A Concrete Co-occurrence Table for [$_{VP}$ give REC PAT]

	Ditransitive	**Not *ditransitive***	**Sum**
give	200	1400	1600
Not *give*	650	147750	148400
Sum	850	149150	150000

The same steps as before would yield the following results:

```
(give.ditr <- matrix(c(200, 650, 1400, 147750), ncol=2, dimnames=list(
   VERB=c("give", "not give"), CONSTRUCTION=c("ditr", "not ditr"))))
##           CONSTRUCTION
## VERB       ditr not ditr
##   give      200     1400
##   not give  650   147750
# MI for give in the ditransitive
log2(200 / (1600*850/150000))
## [1] 4.463284
# G2 for give in the ditransitive
glm(give.ditr ~ c("ditr","not ditr"), family=binomial)$null.deviance
## [1] 926.8206
# OR for give in the ditransitive
(200/1400) / (650/147750)
## [1] 32.47253
```

There is a large number of association measures that can be used (see Pecina 2010) but the three above-mentioned ones probably capture the majority of applications; two other common bidirectional measures are (i) the *t*-score and (ii) the *p*-value of the Fisher-Yates exact (FYE) test.

2.3.3 Applications/discussion

As already indicated, AMs are often used to explore semantic and other characteristics of elements on the basis of other elements they are strongly associated with; examples include near synonymy (of lexical items), many different kinds of syntactic alternations (which involve, in a sense, near synonymy of grammatical constructions). In psycholinguistic studies, associations between elements have been useful as a measure of surprisal – how surprising is linguistic material and what impact does that have on processing speeds? Also, effects of priming have been shown to be correlated with verbs' preferences to occur in certain constructions. In computational linguistics, AMs are often an important step in vector-space semantics because they are used to weigh the co-occurrence of items whose distributional similarity is to be quantified. In research on second language acquisition, studies have explored to what degree learners' knowledge of constructions involves knowledge of which verbs native speakers like to use with which constructions. In other applied linguistics contexts, AMs are also used in key words analyses, which is a variant of collocational studies in which one does not determine which words 'like to occur' with which other words but in which one determines which words 'like to occur' in one corpus as opposed to another one. For instance, words that are (most) key/overrepresented in a corpus of engineering textbooks when compared to a general reference corpus might be words that should be taught (preferably) to engineering students (especially in foreign language learning contexts).

2.3.4 Excursus: frequency and association

There are two important caveats to be mentioned. First and as mentioned above, AMs are computed from frequencies of co-occurrence, but do not take dispersion into consideration. Thus, AMs can be extremely unreliable if the dispersion of the relevant frequencies is not taken into consideration. In the above case, the strong association of *swift* to the word *tuttle* is based on only $^{1}/_{4049}$ files, in which that collocation is tagged as a proper name (of a comet). Thus, any AM result is only as good as the frequencies that

enter into it are 'reliable', to which proper (use of) tagging and checks for underdispersion would contribute.

The second caveat is just as important: Some association measures – those that are related to/derived from significance tests such as G^2, p_{FYE}, chi-squared, *t*, … – react to high association between elements (as they should) *but also* already just to high frequencies of the elements involved. For example, in both the following two tables start and higher.freq, word *w* has the same odds of 1 to 7 (50 vs. 350 and 100 vs. 700) to occur in construction *c*, but the G^2-value of the second table is more than 100% greater than that of the first just because the frequency of word *w* in the second table is twice as high as that of the first (400 in start and 800 in higher.freq):

```
(start <- matrix(c(50, 950, 350, 9998650), ncol=2, dimnames=list(
   WORD=c("W", "Not w"), CONSTRUCTION=c("C", "Not c"))))
##        CONSTRUCTION
## WORD      C   Not c
##   W      50     350
##   Not w 950 9998650
glm(start ~ c("something","other"), family=binomial)$null.deviance # G2 of
start
## [1] 622.2269
(higher.freq <- matrix(c(100, 900, 700, 9998300), ncol=2, dimnames=list(
   WORD=c("W", "Not w"), CONSTRUCTION=c("C", "Not c"))))
##        CONSTRUCTION
## WORD      C   Not c
##   W     100     700
##   Not w 900 9998300
glm(higher.freq ~ c("something","other"), family=binomial)$null.deviance # G2
of higher.freq
## [1] 1249.712
```

The *OR*-measure, on the other hand, returns a result that is more in line with what one might expect from the two identical odds: the *OR*-values are quite close to each other (differing by only about 5.5%):

```
(50/350)  / (950/9998650) # OR of start
## [1] 1503.556
(100/700) / (900/9998300) # OR of higher.freq
## [1] 1587.032
```

But see what happens if we have a table with the same frequency of the word *w* as the first table (400) but a much higher association between *w* and *c* (an odds value of 1 vs. 3) such as the table higher.assoc:

```
(higher.assoc <- matrix(c(100, 900, 300, 9998700), ncol=2, dimnames=list(
   ADJ=c("word w", "Not w"), NOUN=c("construction c", "Not c"))))
##         NOUN
## ADJ      construction c   Not c
##   word w            100     300
##   Not w             900 9998700
```

Now, we can see that

- the G^2-value of higher.assoc is not that much higher than that of higher.freq, meaning the G^2-value does not seem to notice the big association difference (because of the lower frequency of *w* (400 vs. 800);
- the *OR*-value of higher.assoc, however, is much higher than that of higher.freq, meaning the *OR*-value notices the big association difference very well:

```
glm(higher.assoc ~ c("something", "other"), family=binomial)$null.deviance
# G2 of higher.assoc
## [1] 1402.604
(100/300) / (900/9998700) # OR of higher.assoc
## [1] 3703.222
```

The same effect can also surface in really counterintuitive ways. For instance, if one does a keywords analysis on the Clinton/Trump Corpus to identify the words that are (strongly) characteristic of Hillary Clinton's campaign speeches compared to Donald Trump's campaign speeches, one will find the following frequency distributions for the words *hillaryclinton* (as part of the phrase *hillaryclinton.com*) and the word *about*:

Table 8: A Concrete Frequency Table for Hillary Clinton in the Clinton-Trump Corpus

	Clinton corpus	**Trump corpus**	**Sum**
hillaryclinton	26	0	26
other words	117263	445730	562993
Sum	117289	445730	563019

Table 9: A Concrete Frequency Table for about in the Clinton-Trump Corpus

	Clinton corpus	**Trump corpus**	**Sum**
about	579	1386	1965
other words	116710	444344	561054
Sum	117289	445730	563019

Incredibly enough, these completely different distributions (26 vs. 0 and 579 vs. 1386), with their completely different odds ratios (201.5 vs. 1.6!) return nearly exactly the same G^2-values of around 81.66, proving even more that researchers should be extremely cautious in interpreting such values given how G^2 conflates two separate dimensions of information.

2.3.5 The desideratum: unidirectional measures

While the above mentioned measures account for probably 80-90% of all studies, they come with an additional and potentially huge disadvantage, namely that they do not distinguish how much *fast* attracts/repels *car* from how much *car* attracts/repels *fast*. This is more than just a technicality, which is easy to recognize by comparing the following collocations, all of which have very high G^2-values (>189) in the spoken component of the BNC: *according to*, *instead of*, *ipso facto*, *upside down*, *at least*, *de facto*, *for instance*, *in vitro*, *of course*, *bona fide*, *Sinn Fein*.

A little deliberation shows that, although these are all highly attracted collocations, these are actually three groups:

- the first four are cases where $word_1$ attracts $word_2$ but not vice versa: *according* is much more highly predictive of *to* after it than vice versa;

- the next five are cases where $word_2$ attracts $word_1$ but not vice versa: *course* is much more highly predictive of *of* before it than vice versa;
- the last two are cases where $word_1$ is very highly predictive of $word_2$ and vice versa.

Thus, it is usually a good idea to see what a high bidirectional AM reflects, a strong association in maybe just one direction or truly mutual association? One way to compute this is a measure called ΔP, which is the difference between the relevant conditional probabilities of occurrence. ΔP(*give* →ditransitive) is the conditional probability of the ditransitive given *give* minus the conditional probability of the ditransitive given the absence of *give* and is thus computed as follows:

$$\Delta P_{give \rightarrow ditransitive} = \frac{a}{a+b} - \frac{c}{c+d}$$

With our current data in give.ditr, this means we compute it like this in R:

```
give.ditr
##           CONSTRUCTION
## VERB       ditr not ditr
##   give      200     1400
##   not give  650   147750
(200/(200+1400)) - (650/(650+147750))
## [1] 0.1206199
```

On the other hand, ΔP(ditransitive→*give*) is the conditional probability of *give* given the ditransitive minus the conditional probability of *give* given the ditransitive and is thus computed as follows:

$$\Delta P_{ditransitive \rightarrow give} = \frac{a}{a+c} - \frac{b}{b+d}$$

In R:

```
give.ditr
##           CONSTRUCTION
## VERB       ditr not ditr
##   give      200     1400
```

```
##   not give  650   147750
(200/(200+650)) - (1400/(1400+147750))
## [1] 0.2259076
```

The results indicate that the ditransitive attracts *give* much more than *give* attracts the ditransitive – depending on a study's focus and depending on the degree to which a study tries to establish connections to, say, cognitive and/or psycholinguistic factors, these kinds of differences can have theoretically important implications.

In the next section, we turn to general statistical methods in corpus linguistics.

3. General statistical methods used in corpus linguistics

3.1 Descriptive statistics

The simplest kind of statistical method involves descriptive statistics, i.e.

- measures of central tendency, which summarize the distribution of a numeric, ordinal, or categorical variable; frequent examples include the mean (regular or trimmed), the median, and the mode respectively;
- measures of dispersion (now in the statistical sense), which represent the diversity/range of a distribution of a variable and, therefore, also represent how well a measure of central tendency summarizes a variable; frequent examples include the standard deviation, the interquartile range or the median absolute deviation, and normalized entropy respectively; in addition, it can often be useful to indicate standard errors and or confidence intervals;
- measures of correlation, which typically fall into the interval [-1, 1] (or [0, 1]) and in the best of cases indicate how much knowledge of one variable helps predict values of another variable.

(See Gries 2021b: Section 3.1 for discussion of all these measures.) While it is of course impossible to generalize across all existing corpus-linguistic uses of descriptive statistics, it is probably still useful to point out a few pitfalls that users of these statistics should avoid.

First and as mentioned above, one should not provide measures of central tendencies (or frequencies) without a measure of dispersion: a mean, median, or mode need to be accompanied by a fitting measure of dispersion. Second, corpus-linguistic data are often not normally distributed and in all such cases using a median and a median absolute deviation (rather than a mean and a standard deviation) is probably more useful. That also means that other statistical measures that rely on normality need to be used with extreme caution. Third, given the often high degree of messiness and non-linearity in the data, the usual correlation coefficients Pearson's r, Spearman's ρ, and Kendall's τ are often insufficient and used even in contexts where they shouldn't be. Consider Figure 2, which shows the correlation of two variables that are clearly strongly related: If one knows the x-axis value of any point, one can predict the corresponding y-value (range) very well, but all three standard correlations return values that do not reflect this fact well: r=-0.4, ρ=-0.15, and τ=-0.05. However, a more powerful approach to correlations or even just a so-called PRE-measure – a proportional-reduction-of-error measure between 0 and 1 that quantifies how much better the variable on the y-axis can be predicted if one knows the variable on the x-axis – does much better (PRE=0.815).

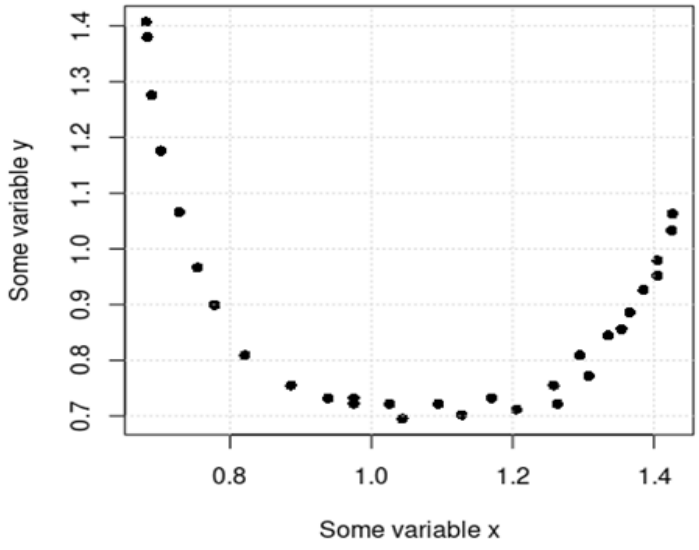


Figure 2: A U-shaped correlation

But such descriptive statistics, while important, are by now usually just the beginning of a real analysis and the next section provides a brief overview of the two most frequent kinds of modeling applications in corpus linguistics.

3.2 Inferential statistics & predictive modeling

3.2.1 Regression modeling

The single most frequent kind of statistical application in corpus linguistics is probably some form of binary logistic regression modeling, i.e. a regression modeling approach that involves a binary response variable (e.g., a phonological, lexical, or constructional choice or the presence/absence of some element) that is predicted on the basis of one or more linguistic and extralinguistic predictors that are suspected to causally influence, or at least be correlated with, speakers' choices. The main advantages of such regression models are that (i) several predictors can be evaluated at the same time, which is more useful and realistic than one multiple single-predictor analyses, and that (ii) interactions between predictors can be included, which means one can determine whether some predictors strengthen, weaken, or even annul the effects of others.

For instance, De Vaere et al. (to appear) study two alternating ditransitive constructions (the indirect- and a prepositional-object constructions) involving German *geben* ('give') in the DeReKo corpus. 1301 occurrences were annotated for 20 morphosyntactic, semantic, and pragmatic factors and submitted to a logistic regression model to see which factors 'make speakers' choose which construction. Interestingly, their regression model involves a variety of more-advanced-than-average methods to obtain good-quality results. For instance, unlike many other studies, they allow for the effect of their numeric predictors to be curved, which is useful because the implicit assumption of straight-line effects is in fact often sub-optimal because it is well-known that many cognitive/psycholinguistic predictors such as learning, forgetting, priming, etc. are curved. Also, they are careful to check their analysis for potential problems such as **overfitting** (inferring too much from the peculiarities of one particular data set) and **collinearity** (the

highly tricky situation where multiple predictors are correlated with each other). They obtain an excellent prediction accuracy (C=0.95) and interpret their findings as providing evidence for the main meaning of *geben* being not so much 'literal transfer from one person to another' (as in *give* or *hand*) but a more general 'transfer' meaning and highlight the descriptive fact that one of the constructions is strongly associated with the passive voice.

Another example for regression modeling is Wulff & Gries (2019), who use a multi-step procedure to study particle placement (e.g. *He picked up the book* vs. *He picked the book up*) in learner corpus data. In a first step, they fit a mixed-effects model to native speaker data to identify factors that co-determine constructional choices in native language. Then, they apply that model to the learner data to determine what native speakers would have said in the situations the learners were in and, ultimately, check to what degree the learners' performance was not nativelike and where. Their study is noteworthy for (i) how their mixed-effects modeling can address at least to some extent effects that are particular to individual particles as well as speakers (in addition to an effect for L1 family) and for (ii) its inclusion of phonological predictors (such as rhythmic and segment alternation) in a study of a morphosyntactic constituent order alternation. As an example of an interaction, they find that the Chinese learners of English seem to pay less attention to the cue of whether a directional PP follows the verb-particle construction than the other learners, who seem to have understood better that *He picked the book up from the floor* is more likely/acceptable than *He picked up the book from the floor*.

Regression modeling will continue to play an important role in corpus linguistics; for many applications, it is probably still the default statistical choice, even if the distributional peculiarities of corpus data – skew/imbalance, noise, Zipfian distribution leading to low cell counts, etc. – often make them challenging to apply; see Hilpert & Blasi (2020), Schäfer (2020), and Gries (2021a, 2021b: Ch. 5-6) for recent overviews/introductions.

3.2.2 Machine-learning / predictive modeling

The probably second most widespread modeling method is the family of tree-based approaches – trees and random forests, see Strobl et al. (2009) – which have emerged as a powerful machine-learning alternative to regression modeling. **Trees** are based on trying to recursively bifurcating 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 reduce the problems of overfitting, collinearity, and overly powerful variables; see Levshina (2020) and Gries (2021b, Ch. 7) for overviews.

An example of a tree-based application is Szmrecsanyi et al., (2016), who study three alternations (genitives, datives, and particle placement) in four varieties of English (Great Britain, Canada, India, and Singapore) to see whether these varieties share a core probabilistic grammar and whether they are split between native and non-native varieties. They use both trees and forests and find that the factors co-determining the alternations indeed have the same kinds of effects (thought different magnitudes) in these varieties, but do not find a neat split between native and non-native varieties.

Corpus linguistics is only slowly warming up to other kinds of machine-learning methods or classifiers. While trees/forests have become more widely used in just the last few years, other methods – neural networks, support vector machines, gradient boosting, to name but a few examples – are more widely used in computational linguistics, but not yet in corpus linguistics, a development that is likely in part due to the sometimes different goals of these two fields: As I see it, corpus linguistics is often more concerned with explaining phenomena (than with pure prediction) and might therefore be more reluctant to adopt more black-boxy methods that excel at prediction but are hard to interpret. However, it is likely that, over time, corpus linguists will explore such methods as well; van der Lee and can den Bosch (2017) is a recent example of a computational-linguistic study that might make corpus linguists see the utility of more diverse classifiers. In their study, they compare multiple classifiers regarding how well they allow to predict the variety or dialect of a language (Netherlandic vs. Flemish variants of Dutch in

a corpus of more than 110,000 subtitle documents) based on text statistics (e.g., average word length or ratio of long/short words), syntactic features (part-of-speech tags), and lexical *n*-grams. They then compare 5 machine-learning algorithms (including random forests and support vector machines). They find that adaptive boosting scores best (in terms of all criteria: precision/recall/*F* and accuracy), but as far as I know, adaptive boosting is an algorithm that has yet to find its way into corpus linguistic applications.

3.3 Exploratory methods

The final group of statistical methods to be discussed is that of exploratory methods, i.e. methods that typically do not test one or more hypotheses but that serve to identify structure (and maybe generate hypotheses) in potentially large and multivariate data sets. The probably most widely used method is that of **hierarchical cluster analysis** (although other clustering methods exist and are used, too). In hierarchical cluster analysis, the algorithm (i) determines how similar each case in one's data is to each other case using a similarity method defined by the user and then (ii) groups together cases into clusters (i.e. groups) that have a high degree of within-cluster similarity and a low degree of between-cluster similarity. An example using clustering on phonetic data is Moisl et al. (2006), who check the Newcastle Electronic Corpus of Tyneside English (63 interviews) for systematic variation of 156 phonetic features and for whether these correlate with social factors. The results show a clear geographical/dialectal north-south divide as well as groupings reflecting speakers' educational status.

An example of clustering on lexico-grammatical data is Gries and Stefanowitsch (2010), who, in one of three case studies, cluster the first verbs in the *into*-causative ([$_{VP}$ V_1 [$_{NP}$ …] *into* [$_{VP}$ V_2*ing*]] as in *He tricked*$_{verb1}$ *him into paying*$_{verb2}$ *a higher price*) in nearly 10,000 examples from newspaper data from The Guardian. They find a coherent cluster structure that reflects the polysemy of this construction and the verbs in it (with different clusters for verbs of trickery, physical force, as well as positive and negative persuasion verbs), a structure that might have been hypothesized, but would not be discoverable without such exploratory methods. Thus, cluster

analysis is a powerful tool but it can be tricky to implement because of its open-ended exploratory nature and the impact that users' methodological choices can have on the results; see Moisl (2015, 2021) for detailed overviews.

Other exploratory methods that are sometimes found are principal components/factor analyses and correspondence analyses, but these are less widespread at this point; Desagulier (2020) provides a good overview of these methods. Finally, there is some growing interest in methods such as **network analysis**: Ellis, Römer, & O'Donnell (2016), who develop semantic networks for the verb-argument constructions they study (e.g., the V about N construction, the V across N construction, etc.), derive a variety of statistics from those (e.g., betweenness and degree centrality, density, and others), and, maybe most interestingly, apply a community-detection algorithm to them to identify a variety of semantically-related coherent groups of verbs in these constructions that shed light on the polysemy of constructions and the prototypical members of semantic groups of constructions. Another example of a network study is Chen's (to appear) structure of the network of Mandarin Chinese space particles in the constructional schema *zai* + NP + space particle in the 10m-words POS-tagged Sinica corpus. Approximately 26K pairs of nouns and particles from these constructions were analyzed with a network approach based on three inputs: (i) collostruction strengths between nouns and particles from a co-varying collexeme analysis, (ii) similarities between the nouns from a word2vec model, and (iii) cosine similarities between the particles. Chen shows that the network exhibits a scale-free structure, meaning that only a few nodes are frequently connected to other units and that most other nodes are relatively unconnected – a striking emergence of the well-known Zipfian distribution of words in constructional slots on the level of a constructional network. Also, the network indicates that experientially and interactionally more prominent particles exhibit a higher degree of local clustering and, thus, more semantic homogeneity.

4. Concluding remarks

In sum, statistical methods are playing an increasingly vital role in corpus linguistics. While not all corpus studies need very sophisticated statistical methods, the increasingly specific hypotheses that are being tested and the need to control noise in corpus data that experimentalists could simply avoid with good and (pseudo-)random experimental designs make it likely that the field will continue on its current trajectory of using ever more advanced methods, which is a good news-bad news kind of situation. The bad news is that this development will continue to pose learnability challenges for all of us – who can claim to stay on top of both corpus linguistics *and* new statistical developments all the time?! But the good news is that, to the extent that we can deal with the learnability challenge, the resulting methodological/statistical diversity also foreshadows exciting discoveries that methods from even just 10 years might not have been able to produce.

References

Adelman, J.S., Brown, G.D.A., & Quesada, J. F. (2006). Contextual diversity, nor word frequency, determines word-naming and lexical decision times. *Psychological Science, 17*(9), 814-823.

Ambridge, B., Theakston, A., Lieven, E.V.M., & Tomasello, M. (2006). The distributed learning effect for children's acquisition of an abstract grammatical construction. *Cognitive Development, 21*(2), 174-193.

Baayen, R.H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon, 5*(3), 436-461.

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.

Bybee, J.L. (2010). *Language, usage, and cognition*. Cambridge: Cambridge University Press.

Chen, A. C.-H. (to appear). Words, constructions and corpora: Network representations of constructional semantics for Mandarin space particles. *Corpus Linguistics and Linguistic Theory*.

Desagulier, G. (2020). Multivariate exploratory approaches. In Paquot, M. & Gries, St.Th. (eds.), *A practical handbook of corpus linguistics* (pp. 435-469). New York & Berlin: Springer.

De Vaere, H., De Cuypere, L., & Willems, K. (to appear). Alternating constructions with ditransitive *geben* in present-day German. *Corpus Linguistics and Linguistic Theory*.

Ellis, N.C., Römer, U., & Brook O'Donnell, M. (2016). Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of Construction Grammar. *Language Learning 66* (Suppl. 1, Language Learning Monograph Series). Hoboken, NJ: John Wiley.

Ellis, N.C., & Simpson-Vlach, R. (2005). An academic formulas list (AFL): extraction, validation, prioritization. Paper presented at Phraseology 2005, Université Catholique Louvain-la-Neuve.

Ellis, N.C., Simpson-Vlach, R., & Maynard, C. (2007). The processing of formulas in native and L2 speakers: psycholinguistic and corpus determinants. Paper presented at the Symposium on Formulaic Language, University of Wisconsin, Milwaukee.

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

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

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. (2015). More (old and new) misunderstandings of collostructional analysis: on Schmid & Küchenhoff (2013). *Cognitive Linguistics, 26*(3), 505-536.

Gries, St.Th. (2020). Analyzing dispersion. In Paquot, M. & Gries, St. Th. (Eds.), *A practical handbook of corpus linguistics* (pp. 99-118). New York & Berlin: Springer.

Gries, St.Th. (2021a). (Generalized linear) Mixed-effects modeling: a learner corpus example. *Language Learning, 71*(3), 757-798.

Gries, St.Th. (2021b). *Statistics for Linguistics with R*. 3rd rev. & ext. ed. Boston & Berlin: De Gruyter.

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

Gries, St.Th., & Stefanowitsch, A. (2010). Cluster analysis and the identification of collexeme classes. In Rice, S. & Newman, J. (Eds.), *Empirical and experimental methods in cognitive/functional research* (pp. 73-90). Stanford: CSLI.

Harris, Z.S. (1970). *Papers in structural and transformational linguistics*. Dordrecht: Reidel.

Hilpert, M., & Blasi, D.E. (2020). Fixed-effects regression modeling. In Paquot, M. & Gries, St.Th. (Eds.), *A practical handbook of corpus linguistics* (pp. 505-533). New York & Berlin: Springer.

Levshina, N. (2020). Conditional inference trees and random forests. In Paquot, M. & Gries, St.Th. (Eds.), *A practical handbook of corpus linguistics* (pp. 611-643). New York & Berlin: Springer.

Lijffijt, J. & Gries, St.Th. (2012). Correction to "Dispersions and adjusted frequencies in corpora". *International Journal of Corpus Linguistics, 17*(1), 147-149.

McDonald, S.A., & Shillcock, R.C. (2001). Rethinking the word frequency effect: the neglected role of distributional information in lexical processing. *Language and Speech, 44*(3), 295-323.

Moisl, H. (2015). *Cluster analysis for corpus linguistics*. Amsterdam & Philadelphia: John Benjamins.

Moisl, H. (2020). Cluster analysis. In Paquot, M. & Gries, St.Th. (Eds.),

A practical handbook of corpus linguistics (pp. 401-434). New York & Berlin: Springer.

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

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

Schäfer, R. (2020). Mixed-effects regression modeling. In Paquot, M. & Gries, St.Th. (Eds.), *A practical handbook of corpus linguistics* (pp. 535-561). New York & Berlin: Springer.

Schmid, H.-J., & H. Küchenhoff. (2013). Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics, 24*(3), 531-577.

Stefanowitsch, A. & St.Th. Gries. (2003). Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics, 8*(2), 209-243.

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.

Szmrecsanyi, B., Grafmiller, J. Heller, B., & Röthlisberger, M. (2016). Around the world in three alternations: Modeling syntactic variation in varieties of English. *English World-Wide, 37*(2), 109-137.

van der Lee, C., & A. van den Bosch. (2017). Exploring lexical and syntactic features for language variety identification. *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, 190-199.

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.

Wulff, S., & Gries, St.Th. (2019). Particle placement in learner English: Measuring effects of context, first language, and individual variation. *Language Learning, 69*(4), 873-910.