13

How to use statistics in quantitative corpus analysis

Stefan Th. Gries

To begin with what sounds like a disappointing start for this chapter: Usually, corpora do not directly provide what most linguists are interested in, such as meaning, communicative function/intention, information structure, cognition/processing and language proficiency/dialect. Instead, a prototypical corpus provides information on the presence or absence of character strings (typically a grapheme such as any letter [in any language], a space, a number or special characters like "%", "~", "@", "™", etc.):

- In (certain parts/locations of) corpora;
- In the presence or absence of other character strings.

Note that character strings can be anything: text that was scraped from the Web, transcribed audio or video data (of spoken or signed language) with or without context and any kind of annotation that was then added to the actual text such as lexical/ structural annotation (on morphemes, parts of speech, parse trees), annotation providing information about the speakers, contextual annotation providing information about the circumstances of language production, etc. That has two consequences: First, whatever a linguist using corpus data is interested in will have to be studied via (i) frequencies of occurrence of something (text or annotation) somewhere in (parts of) a corpus or via (ii) frequencies of co-occurrence of two or more things.

Second, ultimately, corpus-linguistic analysis will involve the notion of *correlation*. For instance, if one wants to study meaning, e.g. the semantics of a certain word or argument structure construction, one typically has to retrieve examples for the word/ construction from one's corpus, annotate them for characteristics of interest and correlate those characteristics with each other and/or with the meanings or kinds of uses of the word/construction. Similarly, if one wants to study information structure and its role for constituent ordering, one typically has to retrieve examples of the constructions in question, annotate them for information-structural and other characteristics and correlate those with each other and the constructional choices.

Trivial as that may seem, this reliance on the notion of correlation also means that corpus linguistics has often a very close to connection to statistical analysis – because

DOI: 10.4324/9780367076399-13

statistics is how we make sense of frequencies, distributions and correlations. This chapter will discuss some of the most frequent kinds of statistical applications in corpus linguistics: I will begin with some statistics that are, in a sense at least, specific to corpus linguistics and that are ordered in terms of how much contextual information they include; I will then turn to applications of statistical methods that are very general, but are here discussed for corpus data annotated for (potentially many) contextual features/ information.

1 Frequency and dispersion

Frequency information

The most basic kind of statistic, and one that is typically completely acontextual, is *token frequency of occurrence*, i.e. the frequency with which something – a morpheme, a word, a multi-word unit, a grammatical construction, etc. – occurs in a corpus or in a part of a corpus; this would lead to statements such as "the word *x* occurs 134 times in corpus *c*". This kind of statement might be compared to another one such as "the word *y* occurs 150 times in corpus *c*", but often we also have to make comparisons between different corpora that are not equally large, in which case we often find normalisations such as "*x* occurs 55 times per million words in corpus *c*, which is more often than *y*'s occurrence of 34 times per million words in corpus *d*". And similar statements can be made for *token frequencies of co-occurrence* such as "in corpus *c*, *criticise* is used 43 times per 100 K passives, but only 13 times per 100 K actives". Such token frequencies have been important in many areas because of their correlations with many experimental tasks (word naming, picture naming, word retrieval); thus, frequency is often used as an explanatory or even just as a control variable in statistical analyses of corpora and experimental data. Also, frequency is a dimension that informs lexicographic work, curriculum/textbook design and many other applications.

Another kind of frequency information is *type frequency*, i.e. the number of types that, for instance, occur in a certain lexically or syntactically defined slot. For example, a corpus might contain 500 instances of the verb lemma *cause* followed by some nominal direct object, and these 500 tokens might instantiate 80 noun lemma types – some very frequent ones (e.g. *problems*, which might account for 120 of the 500 tokens), some intermediately frequent ones (e.g. *pain*, which might account for another 40 tokens) and some really rare ones (e.g. *cerebral palsy*, which might account for just one token). Type frequencies have been considered important for studies of productivity (see Bybee and Thompson 1997; Bybee and Hopper 2001), as when a productive morpheme attached to more types than an unproductive one, or studies of category formation, as when a lexical item becomes a grammatical item by virtue of being associated with many semantically very diverse words. In a sense, very low type frequency can also reveal phraseologisms or fixed expressions: Since the type frequency of the word immediately after the adverb *hermetically* is one in just about every corpus (because *sealed* is just about always the next word), this is a good indicator that it may be a fixed expression.

Given the relevance of frequency information, and especially the relative simplicity with which it can be obtained, frequency data are among the most widely provided corpus statistics. However, frequency is not without its problems, both conceptually and methodologically. For instance, frequency is correlated with many other aspects of language and cognition but that does not entail that it also has a causal effect on these other aspects, which is an entirely different hypothesis to (dis)prove. Statistically and methodologically, however, reporting a frequency, in particular a token frequency, on its own, is just as problematic as reporting an average (such as a mean) on its own: In fact, a frequency of, say, a word *w* in a corpus *c is* kind of a mean, namely the mean of the numbers you get when every word in *c* that is not *w* is coded as 0 and every word in *c* that is *w* is coded as 1. Therefore, the old statistical adage that one should never report an average without a measure of dispersion applies to corpus frequencies, too, and we turn to dispersion now.

Dispersion

The notion of dispersion in corpus linguistics is related to the notion of dispersion in statistics, and it refers to the evenness with which an element is distributed throughout a corpus. From that definition, it also follows that dispersion in corpus linguistics is related to the notion of recency (the effect that we remember and repeat things we have encountered recently more than things we have encountered earlier) in cognitive psychology (see Gries 2019, Chapter 4): What dispersion quantifies is the degree of regularity with which you encounter the element in question when you read the corpus from beginning to end, i.e. the variability in how recently you saw that element if you just saw it again. If a word is distributed very evenly in the corpus, you will see it in regular intervals as you go through the corpus, but if a word is distributed very clumpily (i.e. unevenly), you will not see it at all for a long (corpus) time, then you will see it very often in a very short period of time and then maybe never again.

Consider as an example the words *staining* and *enormous* in the Brown corpus, a corpus containing 500 parts with approximately 2,000 words each of written American English of the 1960s. Both words have the same frequency in the corpus – 37 – but nevertheless their distribution in the corpus could hardly be more different: All 37 occurrences of *staining* are in just 1 of the 500 corpus parts, but the 37 occurrences of *enormous* are spread out over 36 different corpus parts. Put differently, if you randomly pick 1 of the 500 corpus parts, there is only a 0.002 chance it contains *staining*, but there is a 0.072 chance it contains *enormous*. Therefore, trying to characterise both words' distribution by just providing their frequency is completely misleading (and yet still what most corpus studies do).

Of course, this was an extreme example – frequency and dispersion are usually correlated: Obviously, high-frequency function words will be the most evenly dispersed; obviously, it is hapaxes (words with a frequency of 1) that will be most unevenly dispersed. However, in spite of that general correlation, dispersion and frequency deviate from each other most in the range of intermediately frequent lexical words, which is why keeping them separate at all times is essential.

Thus, dispersion should be a central corpus statistic for any corpus-linguistic application involving acquisition/learning, language change or processing, and there are now initial results that indicate dispersion can outperform the predictive power of frequency (Adelman *et al.* 2006; Baayen 2010; Gries 2010). Basically whenever a linguist wants to measure the probability that a linguistic element is known or familiar to a speaker or the familiarity of a linguistic element to a speaker, frequency is usually the easiest statistic to obtain, but should always be augmented by a dispersion statistic; see Gries (2008) and Lijffijt and Gries (2012) for a comprehensive overview of most dispersion measures known then and Gries (2021) for an update of that discussion and an updated script to compute dispersion.

2 Association/contingency

The next kind of basic statistic involves at least a bit of contextual information, namely *co-occurrence information.* In its simplest form, such co-occurrence information may come as frequencies of co-occurrence or probabilities of co-occurrence in the form of statements such as "of the 600 instances of the ditransitive construction (V $\text{NP}_{\text{Recipient}}$ $\text{NP}_{\text{Patient}}$ as in *gave him a book*) in corpus *c*, 200 (i.e. 0.333 or 33.3 per cent) have a form of the verb lemma *give* in its verb slot". Of course, the perspective can also be reversed, as in "of the 400 instances of the verb lemma *give* in corpus *c*, 200 (i.e. 0.5 or 50 per cent) occur in the ditransitive construction". Thus, if one is concerned with the co-occurrence of two elements – *give* and the ditransitive – then their co-occurrence frequency is one number – here, 200 – but to compute a relative frequency/percentage or probability, we need to decide what to normalise it against: the frequency of the construction, i.e. computing *p*(*give*|ditransitive), or the frequency of the verb, i.e. computing *p*(ditransitive|*give*), which amounts to answering different questions. Normalising against the construction's frequency says something about how prominent a role *give* plays for the ditransitive, whereas normalising against the verb's frequency says something about how prominent a role the ditransitive plays for *give*.

However, most applications involving co-occurrence do not just settle for co-occurrence frequency, but also report an association measure, a statistic quantifying the degree of association between the two elements, which is usually computed from a 2 × 2 co-occurrence table such as shown in Table 13.1. In Table 13.1, the italicised frequencies are retrieved with, ideally, a programming language, whereas the regular frequencies are then computed via subtraction from the italicised ones; the letters in the lower-right corners of each cell indicate how these cells are usually referred to in the literature, which means that the most important cell, the one with the co-occurrence frequency, is usually designated *a*.

For such tables, the two relative frequencies mentioned are easy to compute: *p*(*give*| ditransitive) is $^{a}/_{a+c}$ whereas *p*(ditransitive|*give*) is $^{a}/_{a+b}$. However, literally dozens of association measures can be computed from such a table. Two simple ones are *MI* (for *mutual information*) and the *OR* (the *odds ratio*), which are computed as shown in (1) and (2), respectively:

1. $MI = log_2\left(a \div \frac{(a+b)\cdot(a+c)}{a+b+c+d}\right) \approx 6.381$
2. $\text{OR} = \frac{a}{b} \div \frac{c}{d} = 248$ (sometimes, *OR* is reported in its logged version)

If $MI > 0/OR > 1$, then the two elements under consideration occur more often together than expected by random chance (they are "attracted to each other"); if $MI < 0 \,/\, OR < 1$, then they occur less often together than expected by random chance (they "repel each other"); and if $MI = 0/OR = 1$, then they occur together at chance level. In the hypothetical examples provided earlier, therefore, *give* and the ditransitive are strongly attracted to each other. Typically, researchers would compute *MI* or *OR* for all types that might occur in a certain slot/position and then rank-order them. For instance, they

Table 13.1 A co-occurrence table based on the frequencies of co-occurrence (and assuming a corpus size of 100,000 constructions, however defined)

	Ditransitive construction	*Other constructions*	*Sum*
give	*200* *a*	200 *b*	*400* *a* + *b*
other verbs	400 *c*	99,200 *d*	99,600 *c* + *d*
Sum	*600* *a* + *c*	99,400 *b* + *d*	*100,000* *a* + *b* + *c* + *d*

might compute *OR* for every verb type attested in the ditransitive construction and then rank them from the highest *OR* to the lowest.

This kind of statistical method has many applications, virtually all of which are based on the so-called distributional hypothesis:

> If 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. In other words, difference of meaning correlates with difference of distribution.
>
> *(Harris 1970: 785f)*

For our present example, this translates into the expectation that a construction such as the ditransitive will be particularly strongly attracted to verbs whose meanings/functions are compatible with, or even highly similar to, the construction's meaning function. And indeed, in a study of the ditransitive (Stefanowitsch and Gries 2003), *give* and *tell* score the highest association scores, and their semantics are indeed very closely related to the semantics of the ditransitive ("transfer"). Similar considerations apply to the co-occurrence of, for instance, near-synonymous lexical items, whose meaning differences can be so subtle as to be inaccessible even to native speakers, but which can be inferred from other lexical items they are attracted to. For instance, virtually no native speakers are able to explain the difference between *botanic* and *botanical*, yet if one uses association measures to identify the nouns these two forms prefer to modify, clear patterns emerge (see Gries 2003).

However, the range of applications goes well beyond such simple examples: Association measures can be used in, say,

- First-language acquisition contexts: Do children learn syntactic constructions on the basis of the verbs that these constructions are most associated with in their caretakers' speech? How quickly and early do children generalise the use of certain syntactic constructions? (See Pine *et al.* 2013 for a corpus-based study of determiners in English.)
- Second-/foreign-language acquisition/learning contexts: As non-native speakers of a language learn more and more words and constructions, do their preferred usage patterns resemble those of their input or that of native speakers? (See Wulff 2016 for a study of *that* complementation.)

- Diachronic linguistics: Can we quantify the degree of grammaticalisation of verbs by seeing how the degrees of associations with their preferred complements change over time (e.g. by weakening)? (See Hilpert 2006 for discussion of how to monitor linguistic change in diachronic corpus data using association measures.)
- Psycholinguistic studies of speech production: Are speakers more likely to reduce the articulation of words that are highly predictable from, say, the previous word(s)? (See Bell *et al.* 2009 for a study of word durations based using, among other things, conditional probabilities and *MI*.)

In other words, the fact that association measures essentially quantify the notion of contingency (the degree to which two stimuli are probabilistically related), a central component of theories of learning and processing (e.g. Ellis 2006), these measures are probably useful in any scenario in which some linguistic phenomenon might be conditioned or determined at least in part by some other linguistic phenomenon in its context; Evert (2009) provides a good overview of many critical issues. Current topics of discussion involve questions such as "should we use measures that quantify mutual attraction/repulsion (such as *MI*/*OR*) or ones that quantify directional association?" or "should we use measures that are less sensitive to overall frequency/corpus size?" and others, but the general utility of being able to quantify attraction/repulsion of linguistic and other units is hardly ever called into question anymore.

3 Context and concordances

The last somewhat more specifically corpus-linguistic kind of statistic is on a form of output that many corpus linguists would probably not even apply any statistics to: the concordance display, i.e. the display of the search word(s) or tag(s) in question in a central column with a typically user-defined amount of context on the left and right. This is the most informative, context-rich display, because one can see the complete co-text of the expression in question (or, with relevant annotation, the context of the expression). At the same time, it is a display that might appear to defy the very notion of statistical analysis – at least not before annotation has been added, in which case we often apply the methods to be discussed in Sections 4 and 5. However, we will briefly discuss two useful applications of statistical methods to concordances: type-token ratios and lexical gravity.

Type-token ratios

Type-token ratios (TTRs), i.e. the number of types in a certain (part of a) text divided by the number of tokens in the same corpus (part), are a measure of lexical density/richness so it might seem as if this measure could have been mentioned in Section 1. However, since TTRs are very much correlated with the size of a text or a corpus, they are not usually used for that purpose. However, TTRs *can* be more useful when applied to, say, the fixed number of words around a node word or tag of interest. For instance, if one retrieves all instances of two words in a large corpus and then, for every instance of one of the two words, also retrieves exactly 200 words of co-text – 100 before the word in question, 100 after it – then one can compute the TTR for each instance's context and compare them, because then the "text length" in the window has been held constant. What might this be good for? Szmrecsanyi (2006) computes TTRs (on 100 words of context) as a proxy for lexical density in a case study of future choice (*will* vs. *be going to*) in a part of the *British National Corpus* and finds that the TTRs enter into significant and surprising interactions with the main predictor in his study (for instance, he found that increasing TTRs increase the odds for *will*-futures in some corpora, but decrease it in others). Thus, even such a statistically simple method applied to an unlikely target – a concordance display – can yield important results.

Lexical gravity

Another interesting and underused application of statistical tools to concordances is Mason's (1999) notion of lexical gravity. For each slot around, say, a word of interest, he computed the entropy (a measure of randomness) of the frequency distribution of all words in that slot, and the lower the entropy of the slot, the more it deserves attention in the form of an analyst exploring that slot.

How does this work? The entropy H of a frequency distribution is a measure of the evenness of the frequency distribution. In other words, if one has a concordance of a word with 200 instances, then one would look at, say, all 18 slots around the word from nine words to the left (L9) to nine words to the right (R9). Then, for each of these 18 slots, one generates a frequency list of the 200 tokens in it and then computes the corresponding H-value (see Gries 2014: 40–1 for how to compute H). H will be high if the word types making up the 200 tokens are fairly evenly distributed, it will be low if a very small number of types account for most tokens and it will be 0 if a single type accounts for all 200 tokens (recall the example of *hermetically* earlier). Thus, slots for which one obtains a (very) low entropy value will be interesting because it is these slots that the node word whose concordance one is exploring has the strongest impact on. In other words, the entropies function as a pointer towards "where to look next" that is arrived at in a completely data-driven, statistically informed fashion.

4 Regression and classification approaches

The previous sections dealt with, in a sense, specifically corpus-linguistic statistics – "in a sense" because, of course, many disciplines use 2 × 2 co-occurrence tables, but arguably their particular use to compute association measures is a prominent corpus-linguistic method. This section and the next, by contrast, deal with general statistical techniques whose application to corpus data is really not all that different from their use in other areas. The first of these is concerned with regression and classification approaches, where the goal typically is to determine which of potentially very many different (predictor) variables explain speakers' behaviors such as word durations or choices of one of several alternatives, e.g. the choice of one or more functionally similar morphemes, lexical items, grammatical constructions, etc.

In such situations, a researcher's starting point is usually a concordance of a phenomenon in question (e.g. a syntactic alternation between two constructions), which is then imported into spreadsheet software so that each instance can be annotated for all the variables the researcher considers relevant – for instance, because they might be causally related to the phenomenon in question or because they might be variables that only need to be controlled for while one studies the potentially causal predictors. This process typically leads to a spreadsheet or data frame in the so-called *case-by-variable format*: Each instance of the linguistic phenomenon has its own row (i.e. the number of rows corresponds to the sample size *n*), and each variable the data are annotated for has its own column.

Such datasets are then often studied with either a regression-based approach or a tree-based approach. For either method, the researcher formulates one or more hypotheses regarding which (predictor) variables will be correlated with the (response) variable of the phenomenon in question and codifies this hypothesis in a *model*. For instance, if one studied the genitive alternation (*of*-genitive as in *the speech of Captain Picard* vs. *s*-genitive as in *Captain Picard's speech*), one might hypothesise that the choice of one genitive over the other will be related to the length of "the possessor" (*Captain Picard*), the length of the possessum (*the speech*) and the kind or degree of animacy of the possessor (*Captain Picard* is human) and would therefore minimally formulate a model such as that in (3). In this model, the tilde (~) means "as a function of"; therefore, the tilde separates the response variable on the left (the choice of genitive) and the predictors/controls on the right, and in this model the hypothesis is that genitive choices are influenced by possessor length and ("+") possessum length and ("+") possessor animacy.

3. GENITIVE ~ POSSESSORLENGTH + POSSESSUMLENGTH + POSSESSORANIMACY

Both regression and tree-based approaches usually return two types of information: First, they quantify how well the researcher's hypotheses embodied in the model fit the data; second, how much each variable on right of the "~" contributes to the hopefully good fit of the model. The former is often expressed with one of several so-called R^2-values, which range from 0 (very bad fit) to 1 (perfect fit) or other kinds of statistics such as classification or prediction accuracies (how often in a percentage does the model make the right prediction?) and related scores.

The latter, the information about the predictors, usually comes in three kinds:

- An *effect direction*, which states how certain values of the predictors/controls affect genitive choices; for instance, such a model might indicate that the probability of *s*-genitives *in*creases, rather than *de*creases, when the possessor is animate (as opposed to abstract);
- An *effect size*, which states how much certain values of the predictors/controls affect genitive choices; for instance, such a model might indicate that the length of the possessor is a better/stronger predictor of genitive choices than the length of the possessum;
- A *significance test*, which states how likely the effect of a certain predictor/control variable in one's sample would be if, in the population from which the corpus sample was drawn, there was no such effect. If that probability is very small (conventionally below 0.05 or 5 per cent) for a certain predictor, then one typically interprets this as meaning that the effect in one's data is not due to random variation.

With these kinds of information and, typically, some visualisation of the effects that were found, a researcher would then revisit the initial hypotheses: Did the model fit the data well and in a way that confirms the initial hypotheses or not and, hopefully, why is that the case?

While such a model is multifactorial – it considers the potential impact of multiple variables on the phenomenon at the same time – it might still be severely lacking in ways that are still often not understood in the field. This is because the model discussed earlier only tested the so-called *main effects* of the three predictors. For instance, it tested how much possessorlength contributes to genitive choices regardless of the values of the other predictors possessumlength and possessoranimacy, and the same for the other predictors. Thus, the model's results for possessorlength assume and imply that possessorlength has the same effect no matter whether the possessum is short or long and/or whether the possessor is animate, inanimate/concrete or abstract. Thus, if the effect of possessorlength on genitive choices is different for different degrees of possessoranimacy, this model can, by definition, not reveal that because it does not contain a predictor encoding that possibility; that is, by formulating the model as we did in (3), we forced the model to assume that each predictor has the same effect everywhere.

In order to be able to analyse such questions, one's model needs to contain what is called an *interaction term*. An interaction term of predictors A and B, written as A:B, allows the effect of predictor A to vary as a function of another predictor B. To use a non-linguistic example: The effect of taking a certain medication (predictor A) may depend on whether one is taking it with a glass of water or a glass of vodka – if the medication is taken with water, it helps; if it is taken with vodka, it might make matters (much) worse.

In our genitive example, if one expected that the effect of possessorlength was *not* the same regardless of whether the possessor is animate or not, one would include an interaction term in the model, which is often written as in (4):

4. genitive ~ possessorlength + possessumlength + possessoranimacy + possessorlength:possessoranimacy

If the interaction term now returns, say, a significant result with a strong effect, then we interpret this as confirmation that the effect of possessorlength does indeed vary depending on possessoranimacy.

This scenario may seem useful in only a small number of circumstances – but that impression would be mistaken: The notion of interaction is one of the most frequent and important ones for such analyses. First, this is so because we already know that predictors often modify the effect of other predictors. For instance, in the case of particle placement (the alternation between *Captain Picard gave back the phaser* and *Captain Picard gave the phaser back*), there is a strong tendency for idiomatic verb phrases to prefer the particle before the direct object (DO), but when the DO is pronominal, the idiomaticity suddenly "doesn't matter" anymore, and the particle goes behind the DO goes behind the DO. In other words, idiomaticity does not have the same effect everywhere: It has a strong effect when the DO is lexical but none when it is pronominal – that is an interaction.

Interactions play an even more important role in a second way. Imagine the earlier study of the genitive alternation was actually diachronic, covering three different time periods. In that case, one would have another predictor time with three levels. However, including time as a predictor as in (5) is not enough:

5. genitive ~ possessorlength + possessumlength + possessoranimacy + time

This is because this regression model would only indicate whether each of the three linguistic predictors has an effect regardless of what the other two linguistic predictors and time are doing. And from an effect of time, we would only learn whether *s*-genitives become more or less frequent over time, but *not* whether the effects of possessorlength, POSSESSUMLENGTH or POSSESSORANIMACY changed over time – for that we would need each of these variables to interact with TIME, as in (6).

6. GENITIVE ~ POSSESSORLENGTH + POSSESSUMLENGTH + POSSESSORANIMACY + TIME + POSSESSORLENGTH:TIME + POSSESSORLENGTH:TIME + POSSESSORANIMACY:TIME

In other words, the last term, POSSESSORANIMACY:TIME, determines whether the effect of POSSESSORANIMACY has been changing over time: Maybe the effect of animate possessors was stronger 200 years ago than it is now.

The same would be true if one studied the genitive alternation in native and non-native data: To determine whether the three predictors' effects differ between native and non-native speakers, one would need to include (i) a predictor NATIVE with, say, the levels *yes* and *no* and (ii) that predictor's interactions with all other predictors. Or, if one assumed that a speaker's choices change over time in a conversation or in response to what an interlocutor just said (as in priming effects, see Szmrecsanyi 2006 or Hoey 2005), then one most likely needs an interaction of all relevant predictors with a variable that encodes when in a conversation something happened (e.g. a time index or sentence/utterance counter) or what the other speaker just did. In other words, whenever one's thinking about a phenomenon involves the question or expectation that some predictor's effect will not be the same everywhere, one needs a model with interactions of at least that predictor with other variables to do justice to the complexity of one's expectations or hypotheses. The way in which interactions are studied differs between regression models and tree-based approaches (such as classification, regression or conditional inference trees or random forests; see Levshina (2021) for discussion and exemplification), but regardless of how one studies interactions in each approach, it is definitely one of the most fundamental things to consider in one's analyses.

Another important range of issues corpus linguists using these kinds of methods need to be aware of is what one might call the potential *repeated-measurements structure* of one's data: A lot of times, our data contain more than one data point from a speaker. This is important because it means that this speaker's potentially idiosyncratic behavior is represented in the data multiple times, which means that all the data points from that speaker are related, which can heavily distort analyses. Why is that? As a somewhat unrealistic, but nonetheless instructive, example, imagine a learner corpus study of the genitive alternation. We know that the *s*-genitive is very strongly preferred when the possessor is short and animate and the possessum is long and a concrete object. Now imagine a non-native speaker who is at such an early level that they do not know yet an *s*-genitive even exists. That learner might therefore use the *of*-genitive even when everything in the context "screams" *s*-genitive. If that speaker now were to produce multiple such data points, those data points will of course go completely against the overall trend of everyone who actually knows both genitives exist and, thus, weaken the otherwise very robust correlations. Controlling for the fact that these unexpected *of*-genitives all come from the same person is one central reason for the rise of mixed-effects modeling in corpus linguistics. Once one has developed a good general understanding of regression modeling, this kind of approach should be next on one's list.

5 Exploratory analyses

While the previous section was concerned with hypothesis-testing approaches towards regression and classification, another widely used family of approaches are exploratory in nature. In other words, in this kind of work we may have a potentially quite large dataset – again typically cases described by variables in a spreadsheet – and we are interested in identifying any kind of interesting structure in the data, which are too large for "normal human eyeballing" to be successful and reliable. "Any kind of structure" is deliberately general: It can refer to finding (i) groups made up by the cases (i.e. rows) or (ii) groups made up by the variables (i.e. columns), and often the algorithms then identify these groups of cases or variables by trying to maximise within-group similarities and minimise between-group similarities. That means the goal of these algorithms is to create groups that contain members that are as similar as possible to each other while at the same are as different as possible from the members of other groups. For some of these methods, the researcher needs to provide the algorithm with a number of groups to identify; in others, the algorithm will either "propose" a number on the basis of the data that the researcher can then accept or reject, or the algorithm will represent the data in a certain fashion from which the researcher can then pick a most suitable number of groups or dimensions.

The following two sections briefly discuss the probably most frequently used methods: hierarchical cluster analysis followed by principal components analysis and correspondence analysis.

Hierarchical cluster analysis

The point of a hierarchical cluster analysis is usually to find groups among the cases. For instance, we might have different speakers or different languages for which we have numeric or categorical data, and we want them to be grouped, or clustered, on the basis of the data. Moisl (2015), a book-length treatment of cluster analysis for corpus data, gives as an example (Chapter 2, Table 2.3) the application of clustering 24 speakers on the basis of the frequencies with which the speakers used 12 different phonemes in a corpus. Looking at 24 · 12 = 288 numbers is not going to help that much, but for a hierarchical cluster analysis, this is actually a small dataset.

This kind of analysis typically proceeds in three steps. First, the analysis computes a so-called distance matrix, which states for every case how distant it is from – i.e. how dissimilar it is from – every other case. Second, the analysis then computes a cluster structure by successively amalgamating all cases into groups/clusters such that within-group/cluster similarity, or cohesion, is as high as possible. Finally, the resulting structure is represented in a tree with the cases at the bottom and, hopefully, groups/clusters emerging from the connections of the branches. Consider Figure 13.1 for the results of such a cluster analysis applied to nine Russian verbs, all meaning "to try", for which we (Divjak and Gries 2006) had frequency data regarding 87 lexical, morpho-syntactic and semantic features based on 1,585 annotated concordance lines.

Figure 13.1 is instructive in particular because it is not completely obvious how many clusters to assume, an uncertainty that is not uncommon in exploratory analyses. One's first impulse might be to go with three clusters of three verbs each, and that is the solution we adopted; however, at least considering a four-cluster solution might also be useful, namely {*silit'sja*, *poryvat'sja*, *norovit'*}$_1$, {*tuzit'sja*, *tscit'sja*, *pyzit'sja*}$_2$, {*probovat'*}$_3$ and {*pytat'sja*, *starat'sja*}$_4$. The results from this three-cluster solution were then interpreted in terms of a radial network of senses and its relation to the treatment of these near synonyms in traditional Russian lexicography.

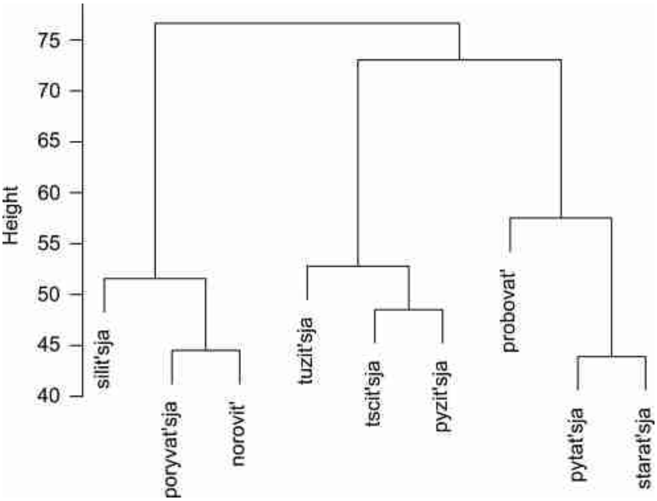

Figure 13.1 Dendrogram of the data discussed in Divjak and Gries (2006)

It needs to be highlighted here that this kind of exploratory approach is not necessarily completely objective: Not only does one have to interpret the cluster structure returned by the analysis, which can be subjective to a certain degree, the analyst also has to make and defend some choices during the process. For example, the computation of the distance matrix can be done with different distance or similarity measures that do not all return identical results; similarly, there are different ways in which the cases can be amalgamated into clusters and, again, the researcher must choose one and justify the choice. On the one hand, this may seem like a weakness of the approach because it seems to indicate a lack of “a clear answer”; on the other hand, we are talking about *exploratory* approaches, so it should not come as a surprise that different ways of exploring the same data can lead to different aspects of the data being highlighted in these different analyses. Gries (2013: Section 5.6) and Desagulier (2018: Section 10.6) are useful first sources to consult, but Moisl (2015) provides a much more comprehensive discussion.

Principal components analysis and (multiple) correspondence analysis

The two analytical methods of this section are conceptually similar. Both take as input matrices of often co-occurrence frequencies (like Table 13.1, just with many more rows and columns) and try to represent the multiple dimensions of information contained in these matrices using much fewer – often just two or three – dimensions; these dimensions are usually orthogonal (i.e. mutually independent or uncorrelated), and both analyses will also indicate how much of the information in the original matrix the now two or three dimensions still represent. Principal component analyses (PCAs) are usually done on numeric measurements, whereas (multiple) correspondence analyses ([M]CAs) are usually done on frequency data.

For instance, a PCA might indicate that a dataset with 20 columns can actually be reduced to, or compressed into, a dataset with only 4 columns (then called principal components) while still retaining 85 per cent of the information contained in the original 20 columns. The PCA would manage that by detecting correlations between the original columns – in other words, redundant information – and then "merging" multiple columns into a principal component. In a somewhat similar vein, a CA decomposes a chi-squared statistic of a co-occurrence table into orthogonal components – which means correspondence analysis uses interrelations between columns (like a PCA) but also between rows (unlike a PCA) – and then represents those in a two- or three-dimensional plot to reveal distributional patterns impossible to see just from the co-occurrence frequencies themselves.

These methods can be extremely useful both on their own for the description and exploration of multidimensional, or multivariate, corpus data, but PCA is also sometimes used to prepare data for regression modeling of the type discussed in the previous section. This is because regressions and tree-based models sometimes have a lot of difficulties dealing with predictors that are highly correlated with each other, a problem referred to as *(multi)collinearity*. This is because if multiple predictors are highly related to each other, then the regression or tree "does not know" to which of them it should attribute what the response variable is doing. If the relevant predictors are numeric, a PCA could be used to reduce the number of correlated predictors – maybe even down to one – so that the subsequent application of a hypothesis-testing model is less jeopardised. As these techniques become more popular in corpus linguistics – in particular (M)CA has become more widely used especially over the last few years – corpus linguists should become more familiar with them; Levshina (2021) and Desagulier (2018: Section 10.2, and 10.4–10.5) are good places to learn more about these methods.

Further reading

Gries, St. Th. (2013) *Statistics for Linguistics with R*, 2nd edn, Berlin and Boston: De Gruyter Mouton. (A still very useful overview of statistical methods, focusing mostly on corpus data and different kinds of regression modeling, but discussing also hierarchical cluster analysis.)

Levshina, N. (2015) *How to Do Linguistics with R: Data Exploration and Statistical Analysis*, Amsterdam: John Benjamins. (A textbook on R in linguistics in general, with many applications pertinent to the sections in this chapter.)

Paquot, M. and Gries, St. Th. (eds) (2021) *Practical Handbook of Corpus Linguistics*, Berlin and New York: Springer. (A new handbook of corpus linguistics with overview chapters on many central corpus-linguistic notions, as well as many hands-on chapters on statistical techniques applied to corpus data using R.)

References

Adelman, J. S., Brown, G. D. A. and Quesada, J. F. (2006) 'Contextual Diversity, Not Word Frequency, Determines Word-Naming and Lexical Decision Times', *Psychological Science* 17(9): 814–23.

Baayen, R. H. (2010) 'Demythologizing the Word Frequency Effect: A Discriminative Learning Perspective', *The Mental Lexicon* 5(3): 436–61.

Bell, A., Brenier, J. M., Gregory, M., Girand, C. and Jurafsky, D. (2009) 'Predictability Effects on Durations of Content and Function Words in Conversational English', *Journal of Memory and Language* 60(1): 92–111.

Bybee, J. L. and Thompson, S. A. (1997) 'Three Frequency Effects in Syntax', *Berkeley Linguistics Society* 23: 65–85.

Bybee, J. L. and Hopper, P. J. (2001) *Frequency and the Emergence of Linguistic Structure*, Amsterdam and Philadelphia: John Benjamins.

Desagulier, G. (2018) *Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics*, Berlin and New York: Springer.

Divjak, D. S. and Gries, St. Th. (2006) 'Ways of Trying in Russian: Clustering Behavioral Profiles', *Corpus Linguistics and Linguistic Theory* 2(1): 23–60.

Ellis, N. C. (2006) 'Language Acquisition as Rational Contingency Learning', *Applied Linguistics* 27(1): 1–24.

Evert, S. (2009) 'Corpora and Collocations', in A. Lüdeling and M. Kytö (eds) *Corpus Linguistics: An International Handbook*, Vol. 2, Berlin and New York: Mouton de Gruyter, pp. 1212–48.

Gries, St. Th. (2003) 'Testing the Sub-Test: A Collocational-Overlap Analysis of English *-ic* and *-ical* Adjectives', *International Journal of Corpus Linguistics* 8(1): 31–61.

Gries, St. Th. (2008) 'Dispersions and Adjusted Frequencies in Corpora', *International Journal of Corpus Linguistics* 13(4): 403–37.

Gries, St. Th. (2010) 'Dispersions and Adjusted Frequencies in Corpora: Further Explorations', in St. Th. Gries, S. Wulff, and M. Davies (eds) *Corpus Linguistic Applications: Current Studies, New Directions*, Amsterdam: Rodopi, pp. 197–212.

Gries, St. Th. (2014) 'Quantitative Corpus Approaches to Linguistic Analysis: Seven or Eight Levels of Resolution and The Lessons They Teach Us', in I. Taavitsainen, M. Kytö, C. Claridge and J. Smith (eds) *Developments in English: Expanding Electronic Evidence*, Cambridge: Cambridge University Press, pp. 29–47.

Gries, St. Th. (2019) *Ten Lectures On Corpus-Linguistic Approaches In Cognitive Linguistics*, Leiden and Boston: Brill.

Gries, St. Th. (2021) 'Analyzing Dispersion', in M. Paquot and St. Th. Gries (eds) *Practical Handbook of Corpus Linguistics*, Berlin and New York: Springer, pp. 99–118.

Harris, Z. S. (1970) *Papers in Structural and Transformational Linguistics*, Dordrecht: Reidel.

Hilpert, M. (2006) 'Distinctive Collexeme Analysis and Diachrony', *Corpus Linguistics and Linguistic Theory* 2(2): 243–56.

Hoey, M. (2005) *Lexical Priming: A New Theory of Words and Language*, London: Routledge.

Levshina, N. (2021) 'Conditional Inference Trees and Random Forests', in M. Paquot and St. Th. Gries (eds) *Practical Handbook of Corpus Linguistics*, Berlin and New York: Springer, pp. 611–43.

Lijffijt, J. and Gries, St. Th. (2012) 'Correction to "Dispersions and Adjusted Frequencies in Corpora"', *International Journal of Corpus Linguistics* 17(1): 147–9.

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*, Amsterdam: Rodopi, pp. 267–80.

Moisl, H. (2015) *Cluster Analysis for Corpus Linguistics*, Boston and Berlin: De Gruyter.

Pine, J. M. Freudenthal, D., Krajewski, G. and Gobet. F. (2013) 'Do Young Children Have Adult-Like Syntactic Categories? Zipf's Law and the Case of the Determiner', *Cognition* 127(3): 345–60.

Stefanowitsch, A. and Gries, St. Th. (2003) 'Collostructions: Investigating the Interaction of Words and Constructions', *International Journal of Corpus Linguistics*, 8(2): 209–43.

Szmrecsanyi, B. (2006) 'Language Users as Creatures Of Habit: A Corpus-Based Analysis of Persistence in Spoken English', *Corpus Linguistics and Linguistic Theory* 1(1): 113–50.

Wulff, S. (2016) 'A Friendly Conspiracy of Input, L1, and Processing Demands: That-Variation in German and Spanish Learner Language', in L. Ortega, A. E. Tyler, H. I. Park and M. Uno (eds) *The Usage-Based Study of Language Learning and Multilingualism*, Georgetown: Georgetown University Press, pp. 115–36.