3

Quantitative designs and statistical techniques

Stefan Th. Gries

1 Introduction

As is well known, corpus linguistics is an inherently distributional discipline: corpora really only contain strings of elements – letters/characters in the typical case of corpora as text files, phonemes, or gestures in the growing segments of auditory or multimodal corpora. That means that analysts can determine their frequency of occurrence, frequency of co-occurrence, or their dispersion/distribution in corpora and analysts have to operationalize whatever they are interested in – meaning, communicative function/intention, speaker proficiency, . . . – in terms of how this will be reflected in such frequencies of (co)-occurrence or dispersions/distributions. From this perspective, it is obvious that knowledge of the discipline involving the analysis of frequencies/distributions – a.k.a. statistics – should form a central component of corpus linguists' methodological knowledge. However, compared to other social sciences (e.g. psychology, communication, sociology, anthropology, . . .) or branches of linguistics (e.g. psycholinguistics, phonetics, sociolinguistics . . .), most of corpus linguistics has paradoxically only begun to develop this methodological awareness. For now, let's assume that corpus-linguistic methods can be categorized in terms of how much context of the occurrence(s) of a linguistic phenomenon they consider into

(i) a group of methods in which the, say, word or pattern under consideration is not studied involving (fine-grained) contextual analysis: if one only wants to know which of the inflectional forms of the verb *give* is most frequent, one does not need to look at the contexts of these verb forms. These methods involve core corpus-linguistic tools such as frequency lists, collocations, dispersions, and statistics computed directly on these.

(ii) a group of methods in which the word or pattern under consideration is studied by means of a detailed analysis of its context. This usually involves the inspection of concordance lines of an element and their annotation for various linguistic and/or contextual features: if one wants to determine when speakers will use the ditransitive (V $NP_{Recipient}$ $NP_{Patient}$) and when the prepositional dative with *to* (N $NP_{Patient}$ $PP_{to\text{-}Recipient}$), one needs to inspect the whole sentence involving these two patterns and their larger contexts to determine, for instance, the lengths of the patient and the recipient, whether the clause denotes transfer or not, etc. Such data are usually analyzed with general statistical tools, i.e. methods that are applied in the same way as they are in psychology, ecology, and so on.

Corpus linguistics needs to "catch up" with regard to both of these groups. With regard to the former, for instance, corpus linguists have used different association measures to quantify, typically, how much two words are attracted to each other or how much a word is attracted to a grammatical pattern, but critical methodological analysis of the commonly used association measures is relatively rare. With regard to the latter, for example, with very few exceptions (such as Biber's multidimensional analysis or Leech, Francis, and Xu's (1994) multivariate exploration of the English genitive alternation) corpus linguistics has only begun to explore more advanced quantitative tools in the last fifteen years or so – compare that to psycholinguistics, which has discussed more advanced linear models and how to deal with subject-specific and lexical item-specific findings at least since Clark (1973).

In this overview, I will discuss statistical tools in corpus linguistics. Section 2 is devoted to the "first group," i.e. statistics directly involving corpus-linguistic tools; Section 3 then turns to the "second group," i.e. statistics that are usually applied to the annotation of concordances. In each section and subsection, I will first discuss some commonly used methods to provide an easier overview of common questions and methods; then I will provide some pointers to more advanced and/or currently under-utilized methods, whose exploration or wider use would benefit the field. Section 4 will conclude with more general comments.

2 Statistics on core corpus-linguistic methods

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

2.1 Frequencies of occurrence

2.1.1 Frequency lists

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

- raw frequencies: *give*'s frequency in the spoken component of the ICE-GB is 297;
- normalized frequencies: *give*'s frequency in the spoken component of the ICE-GB is ≈0.46575 ptw (per thousand words) or ≈465.75 pmw (per million words);
- logged frequencies: the natural log ln of *give*'s frequency in the spoken component of the ICE-GB is ln 297 = 5.693732 (natural logs are computed to the base of e = 2.7182818, and $e^{5.693732}$=297).

Raw frequencies are easiest to interpret within one corpus, normalized frequencies are most useful when frequencies from differently-sized corpora are compared, and logged frequencies are useful because many psycholinguistic manifestations of frequency effects operate on a log scale. For example, if words *a* and *b* occur 1,000 and 100 times in a corpus, *a* will be recognized faster than *b*, but not $^{1000}/100=10$ times as fast but maybe $^{\log 1000}/\log 100=1.5$ times as fast.

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

- learner corpus research, to document potential over-/underuse by learners compared to native speakers;
- language acquisition corpora, to document how children acquire patterns as they increase the number of different verbs (i.e. the type frequency) filling a slot in a particular construction;
- historical linguistics, to document the in-/decrease of use of particular words or constructions over time.

In spite of the straightforwardness of the above, there are still several underutilized methods and desiderata. One is concerned with the fact that words can theoretically have identical type and token frequencies, but may still be very differently distributed. Consider Figure 3.1, which

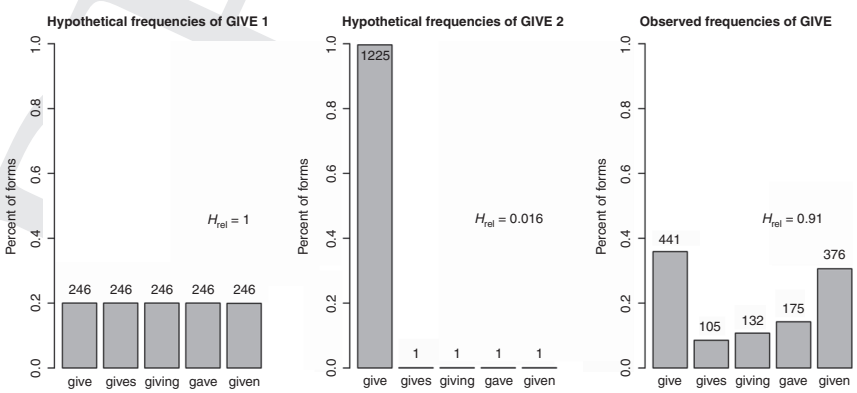

Figure 3.1 Hypothetical and actual frequencies of the forms of GIVE in the ICE-GB and their relative entropies (H_{rel})

shows frequency distributions of the lemma GIVE in the ICE-GB, two hypothetical ones (left and middle panels) and the actual one (from Gries (2010b) in the right panel). While all three distributions have the same token frequency (1,229 instances of GIVE) and type frequency (5 different verb forms), they are obviously very different from each other, which means one should not just report type and token frequencies. One way to quantify these differences is with relative entropy *H*rel as defined in (1) and plotted into Figure 3.1.

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

b. $H_{\text{rel}} = {}^{H}\!/_{H_{max}} = {}^{H}\!/_{\log_2 \text{number of categories}}$

(2) $H_{\text{rel}} \text{ for } \textit{give} = -\left(\frac{441}{1229} \cdot \log_2 \frac{441}{1229} + \ldots + \frac{376}{1229} \cdot \log_2 \frac{376}{1229}\right) \div \log_2 5 \approx 0.91$

Entropies and related information-theoretic measures (e.g. surprisal; see Jaeger and Snider 2008) are not only useful to just descriptively distinguish different frequency distributions as above, but also to questions of language learning or ease of processing in online production.

Even more interesting for frequency lists of words or *n*-grams is the question of what the word or *n*-gram to be counted is or should be. In some corpora one can make use of multi-word unit tags. For example, the *British National Corpus* (BNC) has annotation that shows the corpus compilers considered *of course*, *for example*, *for instance*, *according to*, *irrespective of*, etc. to be one lexical item each, which means one would count *of course*, not *of* and *course* separately. However, in unannotated corpora, the situation is more complicated. Several strategies are possible: first, one can regard spaces and/or other characters as word delimiters and retrieve words or *n*-grams of a particular *n* using these word delimiters. The identification of word delimiter characters is not completely uncontroversial – what does one do with apostrophes, hyphens, etc.? – but far from insurmountable. However, even then the choice of *n* is bound to be arbitrary. To find *according to*, *in spite of*, *on the other hand*, *be that as it may*, and *the fact of the matter is*, one would need to set *n* to 2, 3, 4, 5, and 6 respectively, but typically studies just set *n* to 4 and proceed from there.

A more interesting but unfortunately rarer approach is to let the data decide which *n*-grams to consider. While very useful, these approaches become quite complicated. In one of the first studies to address this problem, Kita *et al.* (1994) proposed to use a cost-reduction criterion, which essentially quantifies how energy (cost) one saves processing a corpus *n*-gram by *n*-gram (where *n* can be any number greater than 0). For each word sequence α, one determines its length in words and its frequency $freq_\alpha$ and len_α in the corpus. From these, one computes the cost

Table 3.1 *The frequencies of several* n*-grams in the untagged Brown corpus*

1-gram	Freq	2-gram	Freq	3-gram	Freq	4-gram	Freq
in	21,428	*in spite*	55	*in spite of*	54	*in spite of all*	3
spite	57	*spite of*	54			*in spite of the*	20
of	36,484	*in ___ of*	625			*in spite of this*	6

reduction $K(a)$ first defined as in (3) and then extended to (4) since word sequences are not mutually disjoint and any shorter n-gram α will be part of a longer n-gram β.

(3) $K(\alpha) = (len_\alpha - 1) \cdot freq_\alpha$

(4) $K(\alpha) = (len_\alpha - 1) \cdot (freq_\alpha - freq_\beta)$ for non-disjoint n-grams such as *in spite* / *in spite of*

Then, all word sequences are sorted by $K(a)$ and the top n elements are considered individual elements of the vocabulary. Finally, one iterates and repeats these steps with the new inventory of individual elements. Consider as an example the n-gram *in spite of* and its parts as well as three 4-grams it is a part of and their frequencies in the Brown Corpus in Table 3.1.

Assuming that *in spite* is a unit is not useful given that, whenever one sees *in spite*, one nearly always also sees *of* as the next word, so the corresponding K-values are very small (see (5); it would be better to assume that *in spite of* is a unit). Correspondingly, assuming that *in spite of* is a unit leads to much higher K-values (see (6)). Thus, this measure quantifies the fact that there is little variation after *in spite*, but a lot more after *in spite of*.

(5) $K(in\ spite(of)) = (2 - 1) \cdot (55 - 54) = 1$

(6) a. $K(in\ spite\ of(all)) = (3 - 1) \cdot (54 - 3) = 102$
b. $K(in\ spite\ of(the)) = (3 - 1) \cdot (54 - 20) = 68$
c. $K(in\ spite\ of(this)) = (3 - 1) \cdot (54 - 6) = 96$

One approach towards the same goal is Gries and Mukherjee's (2010: Section 2.2) implementation of lexical gravity G, which also leads to the notion of lexical stickiness – the degree to which words like to occur in n-grams (cf. Sinclair's 1991 idiom principle) rather than on their own (cf. Sinclair's 1991 open-choice principle). The most sophisticated approaches in corpus linguistics so far, however, seem to be Brook O'Donnell's (2011) "adjusted frequency list" and Wible and Tsao's (2011) hybrid n-grams. The former adjusts frequencies of units on the basis of larger units they occur in (not unlike Kita *et al.*'s work); the latter enriches the study of n-grams with lemma and part-of-speech information (see also the 2010 special issue of *Language Resources and Evaluation* on multi-word units). Other approaches in computational linguistics, which may well inform corpus-linguistic research

Table 3.2 *Damerau's (1993) relative frequency ratio*

	Corpus *T*	Corpus *R*
Perl	249	8
All words	6,065	5,596

in this area, are Nagao and Mori (1994), Ikehara, Shirai, and Uchino (1996), Shimohata, Sugio, and Nagata (1997), and da Silva *et al.* (1999).

2.1.2 Key words

A widespread application of frequency lists is the comparison of frequency lists of two corpora, often one (larger and/or more general) reference corpus *R* and one (smaller and/or more specialized) target corpus *T*. This is useful, for instance, in applied linguistics contexts: if one wants to teach the English of engineering, it would be useful to have a list of words that are more frequent in an engineering context than they are in general English. However, one cannot use a simple frequency list of an English engineering corpus, because its most frequent words would still be *the*, *of*, *in*, . . . – these are frequent everywhere. One of the earliest ways to compare the frequencies of words $w_{1,\ \ldots,\ n}$ in *R* and *T* to determine which words are "key" to *T* compared to *R* involves Damerau's relative frequency ratio. For example, if the word *Perl* occurs in *T* and *R* 249 and 8 times respectively and *T* and *R* contain 6,065 and 5,596 word tokens respectively, then this can be summarized as in Table 3.2. The relative frequency ratio is the odds ratio of this table, i.e. it is computed as $(^{249}/6{,}065) \div (^{8}/5{,}596) \approx 28.72$ and if it is larger/smaller than 1, *Perl* prefers/disprefers to occur in corpus *T* relative to its frequency in corpus *R*. Here we obtain a value much larger than 1, which means *Perl* strongly prefers to occur in *T*.

Another approach towards identifying key words involves G^2, which has been popularized by Dunning (1993) and Scott (1997). For the above data, G^2=270.71, a value indicating very high keyness of *Perl* for corpus *T*.[1]

2.2 Frequencies of co-occurrence

For many linguistic questions, the frequency of occurrence of a word/patterns *P* alone is not sufficient – rather, what is required is the frequency of *P* co-occurring with some other linguistic element *S, T,* . . .. Typically, when *P, S, T*, . . . are words, this co-occurrence is referred to as *collocation* (and *P, S, T*, . . . are *collocates*); when *P* is a construction/pattern, this co-occurrence is referred to as *colligation* or *collostruction*. (and *S*, *T*, . . . are called *collexemes* of *P*). In both cases, a central concern is being able to rank

[1] Many corpus linguists seem to use Paul Rayson's log-likelihood calculator (at http://ucrel.lancs.ac.uk/llwizard.html) but cite Dunning (1993) for the formula, which actually uses a different formula. The above result uses the general G^2 formula in statistics, i.e. the one mentioned by Dunning.

Table 3.3 *Schematic co-occurrence table of token frequencies for association measures*

	S	not *S*	Totals
P	*a*	*B*	*a+b*
not *P*	*c*	*d*	*c+d*
Totals	*a+c*	*b+d*	*a+b+c+d*

collocates/collexemes *S*, *T*, ... in terms of their direction and strength of association with *P*: the words *strong* and *powerful* are near synonyms, but which of them is more likely to be used with *tea* and how much so? Or, the words *alphabetic* and *alphabetical* seem to be very similar semantically, but can we glean how they differ by identifying the words they "like to co-occur with," such as *order* and *literacy*?

More than eighty different measures have been discussed; see Wiechmann (2008) and Pecina (2010) for overviews, and even those do not cover the most recent developments (e.g. Zhang *et al.* 2009 and studies discussed below). Nearly all these measures derive from a 2×2 co-occurrence table such as Table 3.3 (of which Table 3.2 is a reduced version in that it omitted the not-*P* row). If one studied the collocation *alphabetical order*, then (i) *P* could represent *alphabetical, S* could be *order*, and not-*P* and not-*S* would represent all other words, and (ii) the frequency *a* would represent the frequency of *alphabetical order*, which one is interested in, *b* would represent the frequency of *alphabetical* without *order, c* would represent the frequency of *order* without *alphabetical*, and *d* would represent all bigrams with neither *alphabetical* nor *order*.

Typically, association measures involve computing the frequencies one would expect to see in cells *a*–*d* if the distribution in the table followed straightforwardly from the row and column totals (see Gries 2013a: 182). (7) lists a few widely used association measures for the frequencies for *alphabetical order* in the BNC: $a = 87$, $b = 145$, $c = 33{,}559$, and $d = 99{,}966{,}209$. From this, it follows that $a_{\text{expected}} = (87+145)\cdot(87+33{,}559)/(100{,}00{,}000) = 0.078$, etc.

(7) a. $\textit{pointwise Mutual Information} = \log_2 \frac{87}{0.078} \approx 10.12$

b. $z = \frac{a - a_{expected}}{\sqrt{a_{expected}}} \approx 311.11$ and $t = \frac{a - a_{expected}}{\sqrt{a}} \approx 9.32$

c. $G^2 = 2 \cdot \sum_{i=1}^{4} obs \cdot \log \frac{obs}{exp} \approx 1084.84$

It is impossible to single out one association measure as "the best" since they often produce quite different rankings of collocates/collexemes. In the domain of collocation, *Mutual Information* is known to inflate with low

expected frequencies, t is known to prefer more frequent collocations, and G^2 is a quasi-standard. In the domain of collostructions, the $-\log_{10}$ p-value of the Fisher-Yates exact test is used most often (because it is probably the most precise test and a good reference; see Evert 2008: 1235); it remains to be hoped that collocation studies adopt this exact test more.

In spite of the large number of proposed measures, the field still has much to explore. Two areas are particularly noteworthy. The first of these is only concerned with collocations and is concerned with the range of words around a word P that are included. Just as with n-grams, practitioners usually seem to make an arbitrary choice, and frequent choices are 4, 5, or 10 words to the left and to the right, yielding context windows of 8, 10, or 20 words. However, Mason (1997, 1999) has provided a much better solution to this problem, which is unfortunately hardly ever used. He proposes to explore larger contexts of words around P and then for each slot before or after P he computes the entropy of the frequency distribution of the collocates in that slot (along the lines of Section 2.1.1 above). The lower the entropy value, the more a slot deserves attention for the unevenness of its distribution. Table 3.4 exemplifies this approach: the most frequent collocates of *the* in a small corpus are shown in the first column together with their frequencies around *the* in columns 2–4 (3-left, 2-left, 1-left) and 6–8 (1-right, 2-right, 3-right). For example, the circled frequency shows the word *program* occurs 57 times in the position 1-right of *the*. Column 9 shows the entropies of each collocate's frequency distribution and column 10 shows the mean position of the collocate: for *you* and *on*, those are ≈-1, which means these prefer to show up one word before *the*; for *program* and *software*, they are ≈1, which means these prefer to show up one word after *the*. Finally, the last row exemplifies Mason's approach by showing the entropies for the collocate columns (computed on more data than are shown here), and one can see a frequent pattern:

Table 3.4 *Toy example for 3L-3R collocations of* the *with row and column entropies*

Word/Pos	3L	2L	1L	NODE	1R	2R	3R	H_{rel}	Mean
the	9	3	0	194	0	3	9	0.70	0.00
of	0	6	30		0	27	11	0.68	0.61
program	1	2	2		(57)	1	0	0.25	0.79
to	6	9	16		0	6	10	0.86	−0.21
or	4	8	0		0	14	1	0.62	0.11
is	6	3	0		0	12	4	0.69	0.48
and	5	3	3		0	8	5	0.86	0.29
you	4	11	0		0	4	2	0.67	−0.95
on	2	3	13		0	1	1	0.61	−1.00
a	10	0	0		0	2	5	0.52	−0.65
software	1	1	0		6	9	0	0.58	1.12
H_{rel}	0.75	0.75	0.64		0.60	0.67	0.72		

entropies grow with the distance from the node word, which is the technical way of saying that the more slots away one gets from the word of interest, the less systematic patterning one will find.

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

- bi-directional, i.e. they assign a value to, say, the collocation *of course* and do not distinguish whether the association of *of* to *course* is greater/less than that of *course* to *of*;
- based on token frequencies of, again, say, *of* and *course* alone and do not take into account how many different words these two words co-occur with (let alone the entropies of these type frequencies; see Gries 2012a, 2014).

There are two measures, each of which addresses one of these problems, but both need much more exploration and no single measure addresses both problems. As for the former, Ellis (2007) was the first to mention a specifically bi-directional association measure, ΔP from the associative learning literature, in corpus linguistics, which was then used in Ellis and Ferreira-Junior (2009). ΔP is $\approx$0 when no association is present and greater/less than 0 if one word attracts/repels the other (with +1 and –1 being the maximum and minimum values respectively). Consider Table 3.5 with frequency data on *of course* and the two ΔPs (*of* → *course* in (8a) and *course* → *of* in (8b)) as an example.

(8) a. $\Delta P_{course|of} = p(course|of) - p(course|\text{other}) = \frac{5,610}{174,548} - \frac{2,257}{102,35,320} \approx 0.032$

b. $\Delta P_{of|course} = p(of|course) - p(of|\text{other}) = \frac{5,610}{7,867} - \frac{168,938}{104,02,001} \approx 0.697$

Clearly, the word *of* does not attract *course* much – many words can and do occur after *of* – but the word *course* attracts *of* strongly – not many other words occur frequently before *course*. See Michelbacher, Evert, and Schütze (2011) for a discussion of conditional probabilities and ranks of association measures (the latter are promising but come with a huge computational

Table 3.5 *Co-occurrence table for* of *and* course *in the spoken component of the BNC*

	course: present	other	Totals
of: present	5610	168,938	174,548
other	2257	10,233,063	10,235,320
Totals	7867	10,402,001	10,409,898

effort) and Gries (2013b) for a validation of ΔP using multiword units and control 2-grams.

As for the latter problem, lexical gravity *G* (see Daudaravičius and Marcinkevičienė 2004) is an interesting attempt to include type frequencies of collocations in association measures. This measure takes into consideration how many different word types make up a token frequency. Using Table 3.5 as an example again, nearly all association measures would only "note" that there are 2,257 instances of *course* that are not preceded by *of*, but they would not consider how many different words these 2257 tokens represent. The most extreme possibilities are that these 2,257 tokens would be

- 2,257 different word types, which means that *course* was preceded by altogether 1 (*of*) + 2,257 (other) = 2,258 different word types;
- 1 word type only, which means that *course* was preceded by altogether 1 (*of*) + 1 (other) = 2 different word types.

All other things being equal, the first scenario would lead to a higher *G*-value because, anthropomorphically speaking, in both cases *of* managed to sneak into the slot before *course* 5,610 times, but in the first case, it would have managed that although *course* was so promiscuous in terms of allowing many different types in front of it, and this is what *G* would "reward" with a higher value. These and other developments are all in dire need of investigation.

2.3 Dispersion

Another topic that is even more important but at least as understudied is the notion of dispersion, the degree to which any (co-occurrence) frequency of *P* is sensitive to how evenly *P* is distributed in a corpus. For example, if one explores which verbs "like to occur" in the imperative on the basis of the ICE-GB, then many of the most attracted verbs are what one would expect: *let, see, look*, go, *come*, and others – however, two verbs returned as highly attracted stick out: *fold* and *process* (see Stefanowitsch and Gries 2003). Closer inspection reveals that these are fairly frequent in the imperative (esp. given their overall rarity), but occur in the imperative in only a single one of all 500 files of the ICE-GB. Thus, while their association measures suggest *fold* and *process* are strongly attracted to the imperative, their dispersion throughout the corpus suggests that this is such a highly localized phenomenon that it is hardly representative of how *fold* and *process* are used in general.

Ever since some early work in the 1970s (see Gries 2008 for the most comprehensive overview, data for several corpora, and R functions), researchers have attempted to develop (i) dispersion measures that indicate how (un)evenly an item *P* is distributed in a corpus or (ii) adjusted frequencies, i.e. frequencies that are adjusted (downwards) for elements that are unevenly distributed. For instance, both *amnesia* and *properly* occur 51 times in the ICE-GB but one would probably not ascribe the same importance/centrality (e.g. for foreign-language learners) to both: *amnesia* and *properly*

occur in 2 and 47 files of the ICE-GB respectively so adjusted frequencies proposed by Juilland for both are ≈14 and ≈43.5 respectively, which underscores what, here, is intuitively clear: *amnesia* is much more specialized.

Unfortunately, this problem is a very general one: *any* statistic in corpus linguistics is ultimately based on frequencies in parts of corpora, which means that both dispersion and the notion of corpus homogeneity should always be considered potential threats to our studies. Gries (2006) exemplifies (i) how even the simplest of phenomena – frequencies of present perfects – can exhibit large variability across different divisions of a corpus and (ii) how the degree to which speakers' unconscious linguistic choices can be explained can differ hugely between different corpus parts; his recommendation is to always explore and quantify the homogeneity of the corpus for the pertinent phenomenon and at a certain level of granularity.

Given the straightforward logic underlying the notion of dispersion, the huge impact it can have, and the fact that dispersion can correlate as strongly as frequency with experimental data (see Gries 2010c), dispersion and corpus homogeneity should be at the top of the to-do list of research on corpus-linguistic statistics.

3 General statistics

In this section, I will now turn to statistical tools that are often applied to annotation of corpus data, i.e. to data that emerge from the description – linguistic, contextual, or otherwise – of concordance data; Section 3.1 is concerned with confirmatory statistics (and mentions descriptive statistics in passing); Section 3.2 with exploratory statistics.

3.1 Confirmatory/hypothesis-testing statistics

Confirmatory statistics can be classified according to two main characteristics:

- the number of independent variables, or predictors (often, the suspected causes of some observed effect). A design can be *monofactorial*, which means one analyzes the relation between one predictor and one response/effect (see Section 3.1.1), or it can be *multifactorial*, which means one analyzes the relation between two or more predictors and one response/effect (see Section 3.1.2);
- the nature of the dependent variable(s), or effect(s)/response(s), which is usually either *categorical* (e.g. a constructional choice: ditransitive or prepositional dative) or *numeric* (e.g. a reaction times in ms) and which, thus, affects the choice of statistic chosen: categorical responses usually lead to frequencies whereas numeric responses often lead to averages or correlations.

Table 3.6 *The distribution of different types of NPs across subject/non-subject slots (Aarts 1971: table 4.5)*

	Pronouns/names	±Determiner + head	Totals
Subject	5821	928	6749
Non-subject	2193	2577	4770
Totals	8014	3505	11519

3.1.1 Monofactorial statistics

Monofactorial statistical analyses have been relatively frequent in corpus linguistics for quite a while; the most frequent test is probably a chi-squared test for independence, which tests whether an observed distribution is different from a random distribution. Aarts (1971) is a classic early case in point. She studies the distribution of NP types in English clauses to explore, for instance, what kinds of NPs occur in subject slots. As Table 3.6 shows, subject slots prefer structurally lighter NPs: subjects are pronouns/names 86.2 percent of the time ($^{5821}/_{6749} = 0.862$) whereas non-subjects are pronouns/names 46 percent only of the time ($^{2193}/_{4770} = 0.4{,}597$); according to a chi-squared test (see Gries 2013a: section 4.1.2.2), this is extremely unlikely if there is no correlation between subjecthood and NP lightness.

Another well-known application of chi-squared tests is Leech and Fallon's (1992) study of what word frequency differences between the Brown and the LOB corpus might reveal about cultural differences between the USA and the UK. Predating Damerau's relative frequency ratio, they use a difference coefficient and chi-squared tests to identify words that are more/less frequent in AmE/BrE than one would expect if there was no difference between the varieties. As a final example, Mair *et al.* (2002) compare part-of-speech frequencies between the 1960s LOB corpus and its 1990s counterpart FLOB; using G^2 they find that frequencies of nouns increase considerably over time.

Turning to other monofactorial explorations, Schmitt and Redwood (2011) is an example of the use of correlations. They used the Pearson product–moment correlation *r* to address the question of whether English-Language Learners' knowledge of phrasal verbs (numeric scores in tests) is related to the verbs' frequency in the BNC and find a significant positive correlation: on the whole, the more frequent the phrasal verb, the higher the performance of learners. In addition, they use a *t*-test to see whether learners' reception and production scores differ, and they do.[2] Another example from the same domain is Durrant and Schmitt (2009), who compare the use of adjective–noun and noun–noun collocations by learners

[2] Unfortunately, their characterization of their statistical test does not reveal which *t*-test they used. Also, while they *t*-test whether learners perform differently well in productive and receptive tests, they do not test what learner corpus researchers might actually be most interested in: whether corpus frequency has different effects on production and reception, for which multifactorial methods of the kind discussed in Section 3.1.2 would have been required.

with that of native speakers, which were extracted from essays and whose strength was quantified using different association measures. The values of the association meassures were classified into bands respectively so the authors could explore native and non-native speakers' use of collocations of particular strengths with *t*-tests. One kind of result suggests that non-native speakers make greater use of collocations in terms of tokens but not when type variability is considered as well. As a last example, Wiechmann (2008) explores how well corpus-linguistic association measures (on the association of verbs to NP/S complementation patterns) predict the results of eye-tracking experiments. Again using a correlational measure (R^2 s of (quadratic) regression models), he finds that, apart from the theoretically problematic measure of Minimum Sensitivity (see Gries 2012a: 491f.), the association measure of $p_{\text{Fisher-Yates exact test}}$ predicts the experimental data best.

While I will provide more detailed suggestions regarding how statistics in corpus linguistics can generally be improved below, two comments may already be pertinent here. One is that corpus linguists often do not seem to explore in detail whether the assumptions of tests are met. Many common significance tests require particular shapes or properties of the data studied, but usually there is little mention of whether these assumptions were tested let alone met. With observational data, normality especially is very rare, which means that alternative tests (e.g. Kendall's τ or the *U*-test as in Borin and Prütz's 2004 study of *n*-gram type frequencies of native speakers and learners) or more general tests, such as the under-used Kolmogorov–Smirnov test, may often be more appropriate; see Gries (2013a) for discussion of these tests.

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

- they are multifactorial in the above sense: they consider multiple causes for linguistic choices (such as the choice of an *of* vs. an *s*-genitive) into consideration;
- they involve interactions between the linguistic predictors so that one can determine whether a particular predictor (is the possessor of a genitive construction specific or non-specific?) has the same effect regardless of other predictors (is the possessor singular or plural?): maybe specific possessors make it more likely that speakers would produce an *s*-genitive, but only (or especially) when the possessor is also singular . . .;
- they involve interactions between linguistic predictors on the one hand and data-type predictors on the other. Data-type predictors include, for example, L1 (is the speaker a native speaker or a learner of some variety?), REGISTER (which register/genre is a data point from?), TIME

Table 3.7 *Hundt and Smith's (2009) observed frequencies of English present perfects and simple pasts in LOB, FLOB, Brown, and Frown*

	LOB	FLOB	Brown	Frown	Totals
Pres. perfect	4196	4073	3538	3499	15306
Simple past	35821	35276	37223	36250	144570
Totals	40017	39349	40761	39749	159876

(which time period is a data point from?) etc. Including such interactions is necessary if one wants to determine whether the linguistic predictors have the same effect in each L1/variety, in each register, at each time period, etc.: maybe specific possessors make it more likely that speakers would produce an *s*-genitive, but only (or especially) when the speaker is a Chinese learner (as opposed to a German learner or a native speaker) of English . . .

Unfortunately, multifactorial analyses taking all this into consideration, which are usually regression models (see Gries 2013a: ch. 5), are still in the minority. Tono (2004) is a rare exemplary study that takes especially the third characteristic into consideration. Mostly, studies either run no statistics at all and only report observed frequencies, or they run (many) monofactorial statistics on datasets regardless of whether the data are mono- or multifactorial. Table 3.7 represents an example.

Table 3.7 suggests a monofactorial perspective because it seems as if the choice of tense (in the two rows) is dependent on the corpus (the four columns), but the dataset is in fact multifactorial: the frequencies of tenses can depend on the times the corpora represent, the variety, and a potential interaction as shown in a schematic regression equation in (9), in which the tilde means "is a function of."

(9) TENSE ~ VARIETY (AmE vs. BrE) + TIME (1960s vs. 1990s) + VARIETY:TIME

Hundt and Smith (2009: 51) state, among other things, that "[simple pasts] have also decreased over time" but appropriate multifactorial analysis with a binary logistic regression (see the following section) shows that the slight change of frequencies of past tenses is insignificant. In fact, the only significant effect in this data set is VARIETY – there is no diachronic effect of TIME and no interaction of VARIETY with TIME. As for this effect of VARIETY, Hundt and Smith (2009: 51) state that "we are – again – dealing with stable regional variation," which is correct, and the exact result (present perfects are more likely in BrE than in AmE) is represented in Figure 3.2. However, if one calculates effect sizes (see Section 4 below) the effect is so weak (Nagelkerke R^2=0.0017, C=0.524) that it is hardly worth mentioning (and dangerously close to what one might just obtain from variation due to sampling rather than a real varietal difference).

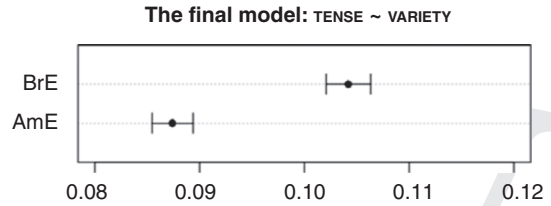

Figure 3.2 The effect of VARIETY on TENSE

Unfortunately, similar examples of multifactorial datasets that are not analyzed multifactorially abound, which is why the recognition that corpus-linguistic statistics has to go multifactorial is maybe *the* most important recommendation for the field's future development.

3.1.2 Multifactorial statistics

Perhaps the most important tool in confirmatory statistics in corpus linguistics is, or should be, the generalized linear model and its extensions, a family of regression models, which serve to model a response/dependent variable as a function of one or more predictors. Crucially, in the GLM and its extensions, the dependent variable can be of different kinds: they can be

- numeric (as when one models, say, numeric test scores as in the above discussion of Schmitt and Redwood 2011), in which case the GLM boils down to "regular" linear regression models;
- ordinal (as when one tries to predict the etymological age of a verb on the basis of characteristics of the verb; see Baayen 2008), in which case one might compute a ordinal logistic regression;
- binary or categorical, in which case one might compute a binary logistic regression (as when above the choice of a tense was modeled on TIME and VARIETY) or a multinomial regression (or a linear discriminant analysis);
- frequency counts (as when one tried to predict how frequency particular disfluencies happen in particular syntactic environments), in which case one might compute a Poisson regression.

In the same way, predictors can also be numeric, ordinal, binary or categorical variables (or any interactions between such variables, see above), and the results of such regressions are predictions (either raw values for linear and Poisson regression or predicted probabilities of outcomes for logistic regressions as in Figure 3.2 and multinomial regressions). The earliest such confirmatory studies that I am aware of – see below for earlier multivariate exploratory methods – are Leech, Francis, and Xu's (1994) use of loglinear analysis to explore the alternation between *of*– and *s*-genitives and Gries's (2000, published 2003a) use of linear discriminant analysis to

study particle placement, the alternation of *John picked up the squirrel* and *John picked the squirrel up*. Following these studies and various replications and extensions – see Gries (2003b) and Kendall, Bresnan, and van Herk (2011) on the dative alternation, Diessel and Tomasello (2005) on particle placement in child language acquisition, Szmrecsanyi (2005) on analytic/synthetic comparatives, particle placement, and future tense, Hinrichs and Szmrecsanyi (2007) on genitives, etc. – such regression analyses have become adopted more frequently, though, see above, not widely enough.

Most of these applications involve binary logistic regressions, i.e. speaker choices of one of two alternatives, but multinomial regression is also slowly becoming more mainstream. Buchstaller (2011) explores the use of multiple quotation markers (*say* vs. *go* vs. *be like* vs. *be all*, and others) in a diachronic corpus of Tyneside speech and finds that the effects of AGE, SOCIALCLASS, TENSE, and NARRATIVE on the choice of quotation marker change over time. Similarly, Han, Arppe, and Newman (forthcoming) model the use of five Shanghainese topic markers on the basis of TOPICLENGTH, TOPICSYNTCAT, GENRE, and other variables. An example for ordinal logistic regression is Onnis and Thiessen (2012), who model levels of syntactic parse depths in English and Korean as a function of *n*-gram frequencies and two conditional probabilities and show, e.g. that cohesive phrases tend to be more frequent and that "the patterns of probability that support syntactic parsing are clearly reversed in the two languages." As a final example, Tono (2004) uses a method that is essentially equivalent to Poisson regressions, namely log-linear analysis, to explore differences between the acquisition of verb subcategorization frames in an EFL context.[3]

While the more widespread adoption of the above tools would already constitute huge progress, there are still a variety of additional improvements that would be useful. First, regressions can be followed up in a variety of ways. One very important one of these is referred to as general linear hypothesis (GLH) tests (see Bretz, Hothorn, and Pestfall 2010). While some scholars now routinely follow Occam's razor and do a regression model selection in which they eliminate insignificant independent variables (as was done above, when, for instance, the interaction VARIETY:TIME was discarded from the discussion of Hundt and Smith's data), what is much rarer is the use of GLH tests to determine whether, say, keeping all levels of a categorical predictor distinct is merited. For example, one might study whether the animacy of a possessor affects the choice of an *s*-genitive and annotate possessors in concordance lines for the following six levels of animacy: abstract vs. concrete/inanimate vs. plants vs. animals vs. superhuman beings vs. humans. However, even if animacy of the possessor were

[3] Interestingly, Tono (2004) changes numeric predictors – target and interlanguage frequencies – into categorical (or, strictly speaking) ordinal factors with three levels (low vs. medium vs. high). This kind of discretization of numeric variables is a not infrequent strategy but, as Baayen (2010) has shown, incurs some loss of power in the statistical analysis. Had Tono done a Poisson regression, this step would not have been necessary; however, this minor issue must not detract from the otherwise very informative statistical analysis.

to play a significant role in the decision for an *s*-genitive, this does not mean that it would be necessary to distinguish all these levels – maybe choices of genitives can be sufficiently well explained even if one just distinguishes two levels: a low-animacy group (conflating abstract, concrete/inanimate, and plants) and a high-animacy group (animals, superhuman beings, and humans). GLH tests can be a very powerful tool to study such questions, discern structure in data, or disprove analyses; see Gries (forthcoming) for a small GLH-based re-analysis of corpus data first discussed by Hasselgård and Johansson (2012).

Second, such regression analyses can be fruitfully combined. Gries and Deshors (2014) develop what they call the *MuPDAR* approach (for *Multifactorial Prediction and Deviation Analysis with Regressions*). This approach is designed to advance learner corpus research and involves three steps and two regressions:

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

Gries and Deshors apply this approach to the use of *may* and *can* by native speakers and French and Chinese learners of English. First, their R_1 determines which factors govern native speakers' use of *may* and *can*. Second, they apply these results to the learner data and predict for each learner use of *may* and *can* which of the two modals a native speaker would have chosen. Third, they explore the cases where the learners did not do what the native speakers would have done to determine what features of the modals the learners still have (the most) difficulties with.

Third, a range of other interesting statistics can help corpus linguistics tackle other statistical challenges. One example is the approach of Structural Equation Modeling, which is designed to help identify causal effects from correlational effects; see Everitt and Hothorn (2011) for an applied introduction. Also, the approach of mixed-effects modeling enjoys a growing popularity in (corpus) linguistics. This method augments the traditional regression methods from above with the ability to include random effects – e.g. subject-, file– or word-specific effects – into the analysis, which has three advantages: (i) it addresses the problem that many statistical techniques assume that the individual data points are independent of each other, which is usually not the case in corpus data where one speaker/writer may provide many concordance examples; (ii) this approach can handle the kind of unbalanced data that corpora provide much better than traditional methods; (iii) since these models help account for, say, subject– or word-specific variability, their results are usually much more precise. Once a variety of uncertainties that still accompany this approach are addressed (see Gries 2013a: 335f.), this will

be one of the most powerful tools in corpus linguistics; see Bresnan *et al.* (2007) for perhaps the inaugural application of this method (to the dative alternation) in corpus linguistics, Baayen (2008: ch. 7 for illustration), and the technique of generalized estimation equations as a potentially more flexible alternative.

Other examples are methods that can help corpus linguists handle the kinds of noisy/skewed data that often violate the assumptions of regression approaches but that are still quite rare in corpus linguistics; examples include classification and regression trees, conditional inference trees, or Random Forests, which, with some simplification involve the construction of flowchart-like tree structures based on successively more fine-grained binary splits of the data; see Hastie, Tibshirani, and Friedman (2009) for technical discussion, Torgo (2011) for more applied discussion, and Bernaisch, Gries, and Mukherjee (2014) for a recent corpus-linguistic application. In addition, the whole field of robust statistics provides a huge array of tools to handle the kind of skewed and outlier-ridden data corpus linguists face very day. Nonlinearities in data may be studied using generalized additive models; see Zuur *et al.* (2009). Finally, the most interesting alternatives to regressions that I have seen in many years are Baayen's (2011) naïve discriminative learning algorithm and Theijssen *et al.*'s use of Bayesian Networks / memory-based learning, both of which have the potential to revolutionize the field in how they provide psycholinguistically more motivated statistics than regression models and allow researchers to build causal models on data that do not meet the usual requirements of regressions (lack of collinearity, for instance).

3.2 Exploratory / hypothesis-generating statistics

Apart from the many confirmatory approaches discussed so far, there is also a large range of so-called exploratory tools, i.e. methods which usually do not test hypotheses and return *p*-values but that detect structure in data that the analyst must then interpret. One of the most widely known methods is of course Biber's multidimensional analysis (MDA); see Biber (1988, 1995) for the most comprehensive treatments. In a nutshell, performing an MDA involves

i. annotating a corpus for a large set of relevant linguistic characteristics;
ii. generating a table of normalized frequency counts of each linguistic feature in each part of the corpus;
iii. computing a factor analysis (FA) on this table, which is a method that will group together those annotated linguistic features that behave similarly in the different corpus parts;
iv. interpreting the co-occurrence patterns in terms of the communicative functions that the co-occurring features perform.

Table 3.8 *Dimensions of variation in Biber (1988)*

Factor	High positive loadings	High negative loadings
1: involved vs. informational production	private verbs, *that* deletion	nouns, long words
2: narrative vs. non-narrative discourse	past tense verbs, third person pronouns	present tense verbs, attribute adjectives
3: situation-dependent vs. elaborated reference	time and place adverbials	*wh*-relative clauses on object and subject positions, pied piping
4: overt expression of argumentation	infinitives, prediction modals	–
5: abstract vs. non-abstract style	conjuncts, agentless passives	–

Biber (1988) identified five dimensions of variation, which are selectively summarized in Table 3.8. MDA has been one of the most influential quantitative methods in corpus linguistics and has spawned a large number of follow-up studies and replications, many of which used MDA results for the characterization of new registers. In addition, MDA has probably been a main reason for why FA, and its statistical sibling, principal component analysis (PCA), have become popular in corpus-linguistic circles long before regression modeling has; see Biber (1993) for an application to word sense identification.

Other exploratory tools that are widespread are cluster-analytic approaches. Just like FA/PCA, cluster analytic approaches try to identify structure in multivariate datasets, but unlike FA/PCA, they do not require the data to be numeric and they return their results in an intuitively interpretable tree-like plot called a dendrogram (see Figure 3.3). Many different kinds of cluster analysis can be distinguished but the most frequent in corpus linguistics is hierarchical agglomerative cluster analysis, which approaches datasets containing n items such that it tries to successively amalgamate the n items into larger and larger clusters until all items form one cluster; it is then the researcher's task to determine how many clusters there are and what, if anything, they reflect. Other techniques are phylogenetic clustering, which is more flexible than hierarchical clustering in that it does not require all elements to form one cluster at some point; k-means clustering, where the analyst defines the desired/suspected number k of clusters, and the analysis returns the n items grouped into k clusters for interpretation; and others.

Given their flexibility, cluster analyses can be and have been applied in very many contexts where large and potentially messy datasets were explored for possibly complex correlational structure that would remain invisible to the naked eye; Moisl (2009) provides a general overview, three recent applications are Divjak and Gries (2006, 2008), who apply cluster analysis to finely-annotated co-occurrence data for nine synonymous

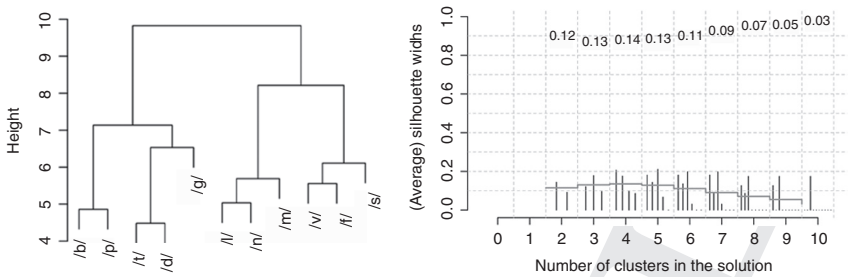

Figure 3.3 Cluster-analytic results for English consonant phonemes (from Gries 2013a)

Russian verbs meaning "to try," Szmrecsanyi and Wolk (2011), who use clustering and other tools within quantitative corpus-based dialectometry, and Hilpert and Gries (2009), who discuss a specific kind of clustering, Variability-based Neighbor Clustering, which can identify temporal stages of development in diachronic corpus data such as longitudinal language acquisition data or historical corpora.

While cluster analysis is not uncommon in contemporary corpus linguistics, there are a variety of follow-up methods that have not been widely adopted yet. These methods can help researchers identify how many clusters to assume for a given dendrogram. For example, it is not immediately obvious how many clusters the dendrogram of English consonant phonemes in the left panel of Figure 3.3 represents: any number between two and five seems possible. The right panel exemplifies one approach to this question, a statistic called average silhouette widths, which quantifies how similar elements are to the clusters which they are in relative to how similar they are to other clusters; in this case, this statistic "recommends" that four clusters should be assumed.

Many more exploratory statistical tools are only used occasionally at this point. Examples include (multiple) correspondence analysis (see Greenacre 2007 for technical details and implementation and Glynn 2010 for an application to the distributional behavior of *bother*) or multidimensional scaling (see Sagi, Kaufmann, and Clark 2011 on how collocates and collocates of *docga*/*dog* and *deo*/*deer* document semantic broadening and narrowing respectively).

4 Discussion and concluding remarks

In spite of having discussed many techniques and desiderata, this chapter could only scratch the surface of quantitative analysis and design in corpus linguistics – most quantitative applications/tools would easily merit an article on their own. As just one example, consider studies concerning the productivity of linguistic elements: Baayen's (1993, 1994) work broke the ground on studies of morphological productivity with applications

ranging from linguistic theory to literary scholarship on vocabulary richness, and Zeldes (2012) is a recent extension of this work to syntax, but lack of space precludes more detailed discussion of these and other works. In addition to the many pointers for future research and exploration above, I will conclude with a few more basic comments and recommendations.

First, in addition to the mere knowledge of what techniques are available, we also need firm guidelines on what is important in statistical analysis, what is important to report, and how methods and results should be reported (see again note 2). Other fields have had long and intense discussions about these things – corpus linguistics, unfortunately, has not. We should be prepared to be inspired by how other disciplines with similar kinds of questions and data have come to grips with these challenges; from my point of view, ecology and psychology are most relevant to us, and Wilkinson and the Task Force on Statistical Inference (1999) provide many essential tips (e.g. to always include effect sizes to distinguish significance from effect size and make analyses comparable).

Let me briefly adduce an example of what happens if even elementary rules of statistics are not followed, in this case the rule that, if one computes many significance tests on one and the same dataset, then one needs to adjust one's significance level (see Gries 2013a: 273f.). The point of this example is not to bash a particular linguist or study – it is only by being able to point out very concrete dangers in existing studies that we learn. The case in point is Egan (2012), who discusses translations of *through* and reports a table (see Table 3.9) in which eight senses of *through* are contrasted in 28 pairwise chi-squared tests.

However, this analysis is quite problematic. One very minor problem is what is presumably just a typo, but since Egan does not provide any actually observed frequencies, there is no way to know whether the comparison between *Channel* and *Means* resulted in 3.4 or 34. Much more importantly, Egan seems to have adopted a critical chi-squared value

Table 3.9 *Chi-squared values with 2* df *for pairwise comparisons (Egan 2012: table 1; figures in italics represent chi-squared values with* p >= *0.05 acc. to Egan)*

	Perc	Space	Channel	Other	Means	Idiom	Time	Clause
Perc		9	6.2	20.9	20.1	54.6	35.5	53.4
Space	9		*0.2*	11	8.8	76	20.3	37.8
Channel	6.2	*0.2*		*2.3*	*3.4*	15.5	6.7	17.3
Other	20.9	11	*2.3*		*5.6*	10	*4.1*	17.6
Means	20.1	8.8	*34*	*5.6*		24	*2.6*	6.6
Idiom	54.6	76	15.5	10	24		17.2	37.6
Time	35.5	20.3	6.7	*4.1*	*2.6*	17.2		7.4
Clause	53.4	37.8	17.3	17.6	6.6	37.6	7.4	

of ≈5.99, the chi-squared value for p=0.05 at df=2. However, Egan did not adjust his critical chi-squared value for the fact that he runs 28 tests on a single dataset. Thus, while he reports 22 significant contrasts out of 28, an adjustment (Hommel's method) results in only 14 significant contrasts, and since it is the "significant" differences upon which his possible network of *through* is based, this network essentially collapses: *perception* senses are not significantly different from all other senses. Similar problems are common: see Gries (forthcoming) for a discussion of a similar flaw in Laufer and Waldman (2011) and Gries (2005b) on how corrections for multiple testing address the issue of too many significant values in keywords analyses. Thus, corpus linguists need to be more aware of the fairly straightforward notion of the multiple-testing problem and ways to address it with corrections and especially corrections that are more powerful than the Bonferroni correction (such as corrections recommended by Holm 1979, Hochberg 1988, or Hommel 1988).

Given all of the above, it may seem as if corpus linguists are supposed to spend quite some time on learning a large number of sometimes quite complex statistical tests. That perception is accurate. As I have asked elsewhere, why is it that corpus linguists look at something (language) that consists of distributional/frequency-based probabilistic data and is just as complex as what psychologists, cognitive scientists, etc. look at, but most of our curricula do not contain even a single course on statistical methods? If we want to make serious headway in our analyses of corpus data, then, given the complexity of our data, we must commit to learning statistical methodology, and hopefully the above succeeded at least in providing an overview of foundational and useful tools that, if adopted, can help us advance our discipline.