3 Quantitative corpus approaches to linguistic analysis: seven or eight levels of resolution and the lessons they teach us

STEFAN TH. GRIES

3.1 Introduction

Over the last fifty or so years, corpus-based methods have developed into one of the most rapidly growing and most widespread 'new' methodology in linguistics. Instead of relying on intuitions of what can or cannot be said, linguists are now turning more and more to corpus data to see what is or is not said.

However, as is only appropriate for a lively scientific discipline, corpus linguistics is still evolving and the field currently witnesses many debates about quite fundamental issues:

1. What is the status of corpus linguistics – is it a theory, a model, a methodology, an approach, etc.?
2. Where does corpus linguistics belong – in the humanities, in the social sciences? In a discourse–analytic context, an applied context, a cognitive/psycholinguistic context?
3. What, if anything, is the difference between corpus-driven and corpus-based approaches and what, if any, implications does this have for our analyses?
4. What is the role of quantitative/statistical work and overall methodological sophistication in the field?

The first two questions have received quite some treatment in a recent special issue of the *International Journal of Corpus Linguistics* (15(3)), and Chapter 2 by Meyer in this volume is concerned with the third question. This chapter, therefore, will deal with the fourth question. While I would never deny qualitative analysis its deserved place in our corpus-linguistic midst, I have argued elsewhere that even very qualitative approaches in corpus linguistics are ultimately based on observing things with particular frequencies (0 or more times), which calls for quantitative analysis (and, of course, quantitative analysis requires interpretation). For example, a statement such as "newspaper coverage of Muslims in British newspapers is increasingly negative" may well be true, as represented in the upper-left panel of Figure 3.1. On the other hand, however,

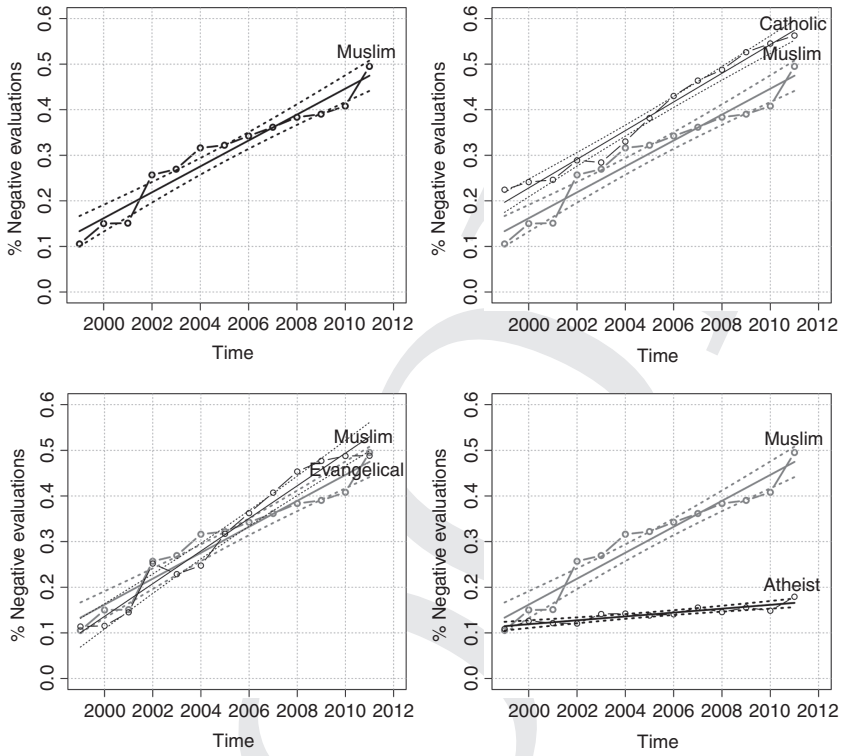

Figure 3.1 A comparison of how much *Muslim*, *Catholic*, *Evangelical*, and *atheist* are used negatively in British journalese (data were made up for expository purposes)

this is probably really only interesting if one can show that the trend that *Muslim* undergoes is different from other words referring to other religious affiliations. In the (made-up) data shown in Figure 3.1, this is sometimes the case and sometimes not. For instance, the upper-right panel shows that the development of *Catholic* is not significantly different from that of *Muslim* ($p \approx 0.17$), the lower-left panel shows that the development of *Evangelical* is significantly different ($p \approx 0.001$), namely steeper, and the lower-right panel shows that the development of *atheist* is also significantly different ($p < 10^{-15}$) but in the other direction, because the negative use of *atheist* has only increased very little.

3.2 What to count, how to count, and why

3.2.1 *The frequency of A in X*

The simplest corpus-linguistic method involves raw frequency counts of any one linguistic feature/expression A in a (part of a) corpus X, as when, for example, in a word frequency list, as schematically exemplified in Table 3.1.

Table 3.1 *Schematic representation of 'the frequency of A in X'*

		Corpus part
A	112	X

While this is no doubt a crude measure, such token frequencies are important in many different research contexts. Among many other things, it has been shown that they correlate with

– the cognitive entrenchment of the referents of words (see Schmid 2000);
– the degree of phonetic reduction and the development of new forms (see Fidelholtz 1975; Schuchardt 1885);
– resistance to regularization in language change (see Bybee and Thompson 1997);
– ease and earliness of acquisition (see Casenhiser and Goldberg 2005);
– subjects' behavior in psycholinguistic experiments such as reaction times in lexical decision, word naming, or picture naming tasks (see Forster and Chambers 1973; Howes and Solomon 1951).

Although frequency counts of A have been useful in all these contexts and more, they are in fact an extremely imprecise measure for a variety of reasons that – while often not discussed – must not be forgotten. One major shortcoming of all frequencies of A in X is their sensitivity to the dispersion of A in X, i.e. the question of how widespread A is in X when different parts of X are studied, where the different parts of X can be linguistically irrelevant (e.g. files in a corpus) or linguistically relevant (e.g. modes, registers, or subregisters).

3.2.1.1 The dispersion of A in linguistically meaningless parts of X

As an example of the distribution of A in a corpus X, consider Leech *et al.*'s (2001) finding that the three words *HIV*, *keeper*, and *lively* occur about equally frequently (≈16 times per million words) in the *British National Corpus* (*BNC*) while, more importantly, they are very differently dispersed throughout the *BNC*: if the *BNC* is divided into 100 equally sized parts, then *HIV* occurs in 62 of these whereas *keeper* and *lively* occur in 97 of these, which corresponds to one's intuition that *HIV* is a word used in a somewhat narrower range of contexts than the latter two.

While the number of parts of a corpus X in which a word A occurs is a valid dispersion measure – sometimes referred to as range – it is a rather coarse measure. Thus, many other measures were proposed, such as Juilland *et al.*'s *D* (0.62, 0.87, and 0.2 for *HIV*, *keeper*, and *lively*), Rosengren's *S*, inverse document frequency, Distributional Consistency *DC*, and many more (see

Gries 2008 for an overview). However, many of these still come with shortcomings:

– some require that the parts of corpus X are equally large, which is unrealistic;
– some are too sensitive (to zeroes or outliers);
– some are too insensitive and return their maximal values (indicating maximally even distributions) too quickly;
– some have ranges of values that don't allow their use for cross-corpus comparison.

A dispersion that does not suffer from such problems is *DP* (see Gries 2008, 2010c). For a word A in a corpus X it is computed as follows:

1. compute the size of each part of X (in % of all of X);
2. compute the relative frequency of A in each part of X;
3. compute the absolute pairwise differences between the sizes and the relative frequencies, sum them, and divide the sum by two.

DP is close to 0 when A is distributed evenly, and close to 1 when A is distributed unevenly/clumpily. Figure 3.2 shows the relation between frequency and *DP* on the basis of words from different frequency bins of the BNC sampler. On the one hand, there is obviously a probabilistic relation between the frequencies of elements and their dispersion, as indicated by the non-parametric smoothers in both panels: the more frequent a word, the more evenly distributed it is throughout the corpus. On the other hand, it is also clear that the correlation between frequency and *DP* is only probabilistic. Especially the middle frequency range contains words with very high and very low dispersions, and the right panel exemplifies this on the basis of sixty-eight random words from five frequency bins. (The words are jittered along the *x*-axis.) It is plain to see that

– words such as *ft*, *diamond*, *Russians*, and *Egypt* are as frequent, but much less evenly distributed than *hardly*, *properly*, and *anywhere*;
– the word *er* is approximately as frequent as *do*, *have*, *be*, and *but*, but much less evenly distributed (because it nearly only occurs in speaking); in fact, *er* is clumpier than all the sampled words that are one order of magnitude less frequent, but about as distributed as *hardly*, which is two orders of magnitude less frequent.

Crucially, this is not just corpus-linguistic playing with numbers. For example, Ellis and Simpson-Vlach (2005) and Ellis (2007) show that even a dispersion measure as crude as range can have significant predictive power above and beyond frequency, and Gries (2010c) shows that some dispersion measures correlate more highly with response time latencies from Balota and Spieler (1998) as well as from Baayen (2008).

The first lesson to be learned therefore is the following:

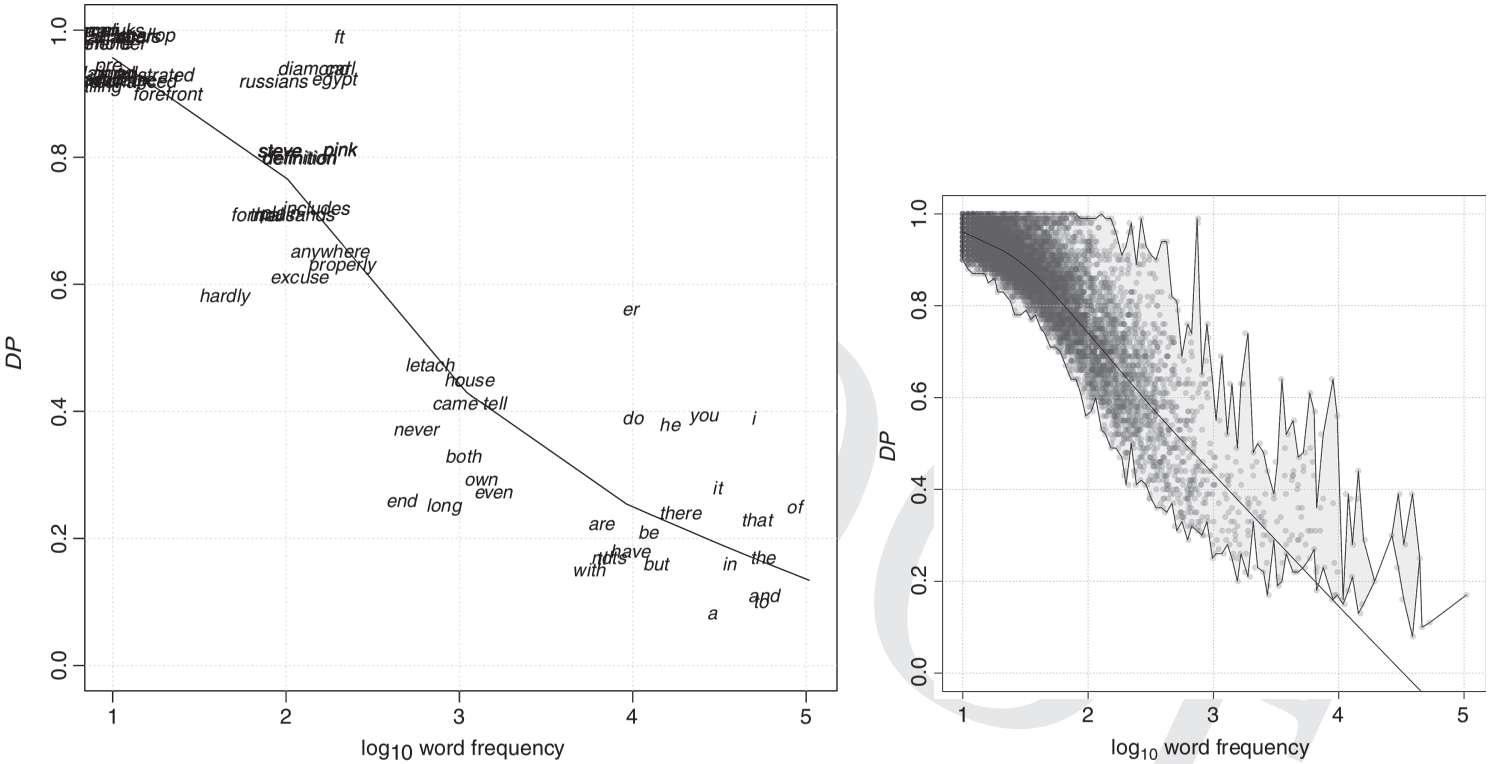

Figure 3.2 The relation between (logged) frequency (on the x-axes) and *DP* (on the y-axes): all words in the BNC sampler with a frequency ≥10 (left panel), 68 words from different frequency bins (right panel)

Lesson 1: Frequencies should always be augmented by, or checked against, dispersion measures.

3.2.1.2 The dispersion of A in linguistically meaningful parts of X

While there are a few – too few – applications of dispersion measures, many of these are based on divisions of the corpus X into parts that are linguistically irrelevant, such as files. However, many of the general corpora most widely in use usually come with a linguistically meaningful structure, namely a division into modes, registers/genres, sub-registers, and so on. While this is good in terms of representativeness, it also means that any statement about what happens (how often) in a corpus X is only a generalization over the parts of X that usually implicitly just assumes that what happens in the parts of X is not (significantly or substantially) different from what happens in X as a whole. In other words, a statement about X relies on the null hypothesis that all parts of X behave as X does as a whole, clearly a rather bold assumption given how heterogeneous corpus parts can be (see below for more on this and Gries 2006 for a proposal on how to measure the homogeneity of a corpus with regard to a particular phenomenon and level of resolution).

For example, in Gries (2006) I show how a phenomenon A as mundane as present perfect frequencies in the British Component of the *International Corpus of English* (ICE-GB) exhibits significant distributional differences on every level of corpus organization:

– present perfects are significantly more frequent in speaking than in writing;
– present perfects are significantly less frequent in printed writing than in non-printed writing and in the spoken registers of dialog and monolog.

In fact, it turns out that the heterogeneity of present perfects within writing but between printed and non-printed writing is larger than that between speaking and writing although the latter is one of the pet distinctions of most corpus studies that many studies consider and although the former one is not. The second lesson to be learned therefore is:

Lesson 2: Frequency effects should always be checked on multiple levels of corpus granularity to explore the homogeneity of the corpus.

In addition to these aspects, there is also an even more general thing to be considered, however. There is a growing body of research that shows that raw decontextualized frequency counts are not as relevant anyway. For instance, Raymond and Brown (forthcoming) find that reduction effects are less due to overall frequency and more due to cumulative exposure and contextual predictability. More radically, Baayen (2010) suggests that frequency effects might be epiphenomenal:

Table 3.2 *Schematic representation of 'the frequency of A in P in X'*

	Context P	Corpus part
A	66	X

word frequency is correlated with many other lexical properties; in fact ... most of the variance in lexical space is carried by a principal component on which contextual measures load highest: syntactic family size, syntactic entropy (!), BNC dispersion (!), morphological family size, adjective relative entropy, variety of contexts) – by contrast ... frequency only explains a modest proportion of lexical variability.

Thus, a move towards more contextualized approaches is definitely desirable, which I will begin to discuss as of the next section.

3.2.2 The frequency of A in P in X

A corpus-linguistic method that takes contexts of A (in X) more into consideration involves the notions of *collocation* and/or *colligation*, where, say, a word is studied in its lexical or grammatical/textual context. Schematically, Table 3.1 changes to Table 3.2.

Again, frequencies of occurrence of something in a particular context have many useful applications and implications. For instance, they correlate with phonological reduction phenomena, grammaticalization, the emergence of prefabricated expressions (see Bybee 2010 for an overview) or the existence of verb islands in first language acquisition (see Tomasello 2005). Nevertheless, the question remains how the occurrence of A in some context P (in a corpus X) is quantified best.

Corpus linguistic studies or studies that use corpus data over the last few decades have adopted two kinds of approaches to studying such context-bound co-occurrence frequencies. On the one hand, and this is the topic of this section, they have used observed co-occurrence frequencies or conditional probabilities. On the other hand, they have used association measures (such as Mutual Information *MI*, *t*, Log-likelihood, $p_{\text{Fisher–Yates exact}}$ ...). However, in a recent publication, Bybee (2010) argues vehemently against the latter approach of association measures in general and one particular approach in particular – collostructional analysis (see Gries and Stefanowitsch 2004a, 2004b; Stefanowitsch and Gries 2003, 2005) – and in favor of the former approach of co-occurrence frequencies/conditional probabilities. Her claims relevant in the current context are as follows (see Gries 2013 for a more comprehensive discussion):

1. lexemes do not occur in corpora by pure chance" (Bybee 2010: 97), which is why *p*-value-based association measures are problematic;
2. the frequency of lexemes in other uses [i.e., other than in the context P under consideration, STG] is not important (Bybee 2010: 100);
3. [s]ince no semantic considerations go into the analysis, it seems plausible that no semantic analysis can emerge from it (Bybee 2010: 98).

From a corpus-linguistic perspective, these are strange statements, to put it mildly. As for 1, no one in his right mind would ever assume that lexemes occur in corpora by pure chance. Of course they do not – if they did, what sense would it make to study co-occurrence data? The motivation to go beyond mere token frequencies or conditional probabilities is to separate the wheat (linguistically revealing co-occurrence data) from the chaff (the fact that nouns co-occur with *the* a lot), and association measures of all kinds – descriptive measures, *p*-values ... – are nearly always only used to rank collocates/collexemes, which downplays the role of significance testing.

In that connection and as for 2, there is in fact evidence that the information above and beyond the mere co-occurrence frequency is useful. For example, experiments on the psychology of learning have shown that association measures such as ΔP, which do involve more than mere co-occurrence are useful to predict subjects' behavior and performance and are in fact highly correlated with, for example, some corpus-linguistic measures such as $p_{\text{Fisher–Yates exact}}$ (see Ellis and Ferreira-Junior 2009). In fact, Bybee contradicts herself: how can she approvingly quote Goldberg (2006) with regard to how "in category learning in general a centred, or low variance, category is easier to learn" (2010: 89), which entails that different types of a category and its token frequencies are relevant, and at the same time insist that the distribution of a word w outside of the construction c is irrelevant? Even more pertinently, Gries *et al.* (2005, 2010) pit co-occurrence frequency, conditional probability, and $p_{\text{Fisher–Yates exact}}$ against each other in a sentence-completion experiment and a self-paced reading-time study, and in both $p_{\text{Fisher–Yates exact}}$ outperforms the competing measures by a wide margin. Similarly, Colleman and Bernolet (2012) find seemingly erratic verb-specific preferences in the Dutch dative alternation that fall neatly into place when explored with an association measure.

As for 3, this claim is absurd, given the large amount of work in computational (psycho)linguistics in which purely frequency-based distributional analyses reveal functionally highly coherent clusters. Two classics are Redington *et al.* (1998) and Mintz *et al.* (2002), who both show how distributional analyses of co-occurrence frequencies reveal clusters that resemble something that, in cognitive linguistics, is considered to have semantic import, namely parts of speech. Even if one did not postulate a relation between parts of speech and semantics, the analysis reveals that something can emerge from a statistical analysis (parts of speech) that did

Table 3.3 *Schematic representation of 'the frequencies of A in P and ¬P in X'*

	Context P	Context ¬ P		Corpus part
A	66	(*a*) 23	(*b*)	X
¬ A	400	(*c*) 500	(*d*)	X

not enter into the analysis (since only bigram frequencies entered into the statistics) and that is a strength of exactly the type of usage-/exemplar-based models that both Bybee and I favor. Even with regard to collostructional analysis, Gries and Stefanowitsch (2010) show that purely statistical analyses of the *into*-causative (NP_{SU} V NP_{DO} *into* V-*ing*) and the *way*-construction (NP_{SUi} V $POSS_i$ *way* PP_{LOC}) result in clearly and finely delimited verb classes.

Given these arguments in favor of association measures, lesson 3 is the following, which will be taken up in more detail in the next section.

Lesson 3: Do not just rely on co-occurrence frequencies of A in P in X, but also include the uses of A elsewhere.

3.2.3 The frequencies of A in P and ¬P in X

The simplest way to include more contexts is the one that has been used most in corpus linguistics: in corpus X, one considers the distribution of A and ¬A in context P and in other contexts (¬P), as exemplified schematically in Table 3.3; the parenthesized letters name the four cells.

As mentioned in the previous section, data like these are usually evaluated with one or more of a range of association measures that have been proposed; recent overviews by Wiechmann (2008) and Pecina (2009) discuss 47 and 82 measures respectively. It seems as if the following measures are most widely used:

– (Pointwise) Mutual Information, which overemphasizes rare events;[1]
– the *t*-score, which is more strongly correlated with overall frequency;
– the log-likelihood ratio, which is probably the closest asymptotic approximation to the Fisher–Yates exact test (see Evert 2009: 1235);
– the (logged) *p*-value of a Fisher–Yates exact test.

While the last of these measures is computationally the most demanding one, it seems to be the one that does most justice to the data in terms of its mathematical characteristics and requirements: as an exact test based on the

[1] See Zhang *et al.* (2009) for a recent modification of *MI* – *EMI* – and its applicability to multi-word units; a function for R to compute *EMI* for 2x2 tables is available from the author upon request, as are functions for dispersion measures, ΔP, and entropies.

hypergeometric distribution, it comes with no assumptions (e.g. normality) and can handle the type of Zipfian data so typical of corpus data. Such measures have been very widely used in semantic/lexicographic applications (identifying strong/significant collocates) and research on argument structure constructions (in the collostructional approach such measures return, e.g. the verbs reflecting the central senses of constructions; such measures are often well-correlated with priming effects and predict phonetic reduction, etc.

Even if one recognizes that this approach is superior to frequency counts alone, the range of measures that have been proposed raises the natural question which of these measures is best. While there may not be one measure that does well in all applications, some evidence at least is mounting that also suggests that $p_{\text{Fisher–Yates exact}}$ is again among the top choices. For example, Wiechmann (2008) finds that its correlation with psycholinguistic data is the second best of all measures tested, and Ellis and Ferreira-Junior (2009) also use $p_{\text{Fisher–Yates exact}}$ and find it to correlate very nicely with frequency of learner uptake.

In an attempt to improve the existing range of association measures, some recent studies have pursued two different approaches to make them more precise. One approach is based on the recognition that virtually all association measures are bidirectional/symmetric, whereas association may be unidirectional. This point was made repeatedly by some scholars (Kjellmer 1991; Smadja 1993; Stubbs 2001; Bartsch 2004; Evert 2009) but so far no attempts have been made to develop measures that can handle this. Ellis (2007) provides some insightful discussion of the role of associative learning for language acquisition and mentions a measure called ΔP, which allows one, in the example of Table 3.3, to quantify not just the association between A and P, but more precisely the association of P to A and the association of A to P:

$$\Delta P_{A|P} = p(A|P) - p(A|not\ P) = \frac{a}{a+c} - \frac{b}{b+d} \quad (1)$$

$$\Delta P_{P|A} = p(A|P) - p(A|not\ P) = \frac{a}{a+b} - \frac{c}{c+d} \quad (2)$$

This is more important than it may seem because Gries (2013) shows that more than 25 percent of multi-word units (as defined by the multi-word tagging of the BNC) are *not* mutually attracted to each other – they really only exhibit a strong association in one direction. For example, in the bigram *of course*, the association *of* → *course* is quite low (0.032) but it is the association *course* → *of* that makes this a multi-word unit (0.697). However, none of the traditional measures reflect this – they only assign A a strong/significant-collocation status, failing to account for its directionality. Similar cases abound.

– $word_1 \rightarrow word_2$ bigrams: *instead of*, *according to*, *owing to*, *pertaining to*, *volte face*, *kung fu*, *gung ho*, *faux pas*
– $word_2 \rightarrow word_1$ bigrams: *for instance*, *for example*, *old-fashioned*, *status quo*, *pot pourri*, *coup d'etat*, *grand prix*

Thus, we arrive at lesson 4:

Lesson 4: Do not just consider attraction between elements to be bidirectional but use ways of investigating things that address the fact that it often is not.

3.2.4 Excursus: In "A in P," what is A anyway?

It is at this point (at the latest) that it is both necessary and instructive to briefly interrupt the progression from simple (co-)occurrence frequency to, ultimately, multidimensional co-occurrence frequencies and more and consider what the *A* is whose (co-)occurrence frequency is being discussed. It is probably fair to say that, typically, in corpus linguistics A corresponds to a word. However, it should be clear that this is a shortcut that is typically just as crude as it is convenient. Take verbs, for instance. It is clear that the context in which a verb is used, or the pattern or construction it is used in, is correlated with its sense: *to run* used intransitively is more likely to refer to the 'fast pedestrian motion' sense and less likely to refer to the 'manage/oversee' sense than *to run* used transitively; similarly, *recognize* when followed by a subordinate *that*-clause means something different from *recognize* when followed by a direct object ('recall to mind' vs. 'acknowledge the truth of,' as was discussed in a recent case before the Supreme Court of the United States). It is also clear that speakers/listeners have access to this information: online comprehension and the resolution of ambiguities by listeners is informed by the frequencies with which verbs occur with/in particular patterns/constructions (see Garnsey *et al.* 1997; Hare *et al.* 2003; Trueswell *et al.* 1993).

While the above has been recognized in psycholinguistics and computational linguistics (see in particular Roland 2001 and other studies by him and collaborators), as far as I can tell, the vast majority of studies in corpus linguistics is still word-based, not sense-based. There are some studies bucking this overall tendency. For example, in the second study on collostructions, Gries and Stefanowitsch (2004a), the authors discuss how the different senses of a verb such as *have on* may exhibit different preferences for constructions. For example, Bernolet and Colleman (2012) show that senses of verbs in Dutch ditransitives and prepositional datives do a better job at explaining subjects' behavior in a priming experiment than the verbs alone. Given all this, lesson 5 is probably all too obvious:

Lesson 5: Do not just consider between-lexeme differences (i.e. differences between different words) but also within-lexeme sense differences.

Table 3.4 *Schematic representation of 'the frequencies of A in P (Q, R, S . . .) in X'*

	Context P	Context Q	Context R	Context S	Corpus part
A	66	23	16	7	X
¬ A	–	–	–	–	X

The next quantitative approach is concerned with the fact that the frequencies *b* and *c* in a contingency table are merely very crude approximations to the real data since all the variability of the words/senses[2] ¬A and all the variability that can be observed in ¬P are forced into one number. The next section will consider with how this issue can be dealt with more appropriately.

3.2.5 The frequencies of A in P (Q, R, S . . .) in X

Given Section 3.2.5, the next more precise step is somewhat obvious and involves looking at A's distribution across its whole range of usage types, which is exemplified in Table 3.4.

In essence, this is just a different type of dispersion: not across files or corpus parts, but across co-occurrence patterns. But what is this relevant for? On the one hand, it paints a descriptively more accurate picture of A's patterning, which, if A is a verb, could be A's subcategorization patterns/preferences or the functions to which A is preferably put to use (see Roland *et al.* 2007). On the other hand, it also has more theoretical implications because data like this allow to again connect more to work in more cognitively oriented work. A table such as Table 4 helps identify the reliability of a form–function cue, which has profound implications for, say, language acquisition studies.

For instance, Casenhiser and Goldberg (2005) tested children between the ages of 5 and 7 to determine whether they learn a new construction better from skewed or from balanced input. Specifically, these two input conditions both involved input in the form of 5 verb types and 16 tokens, but they differed in their type-token distributions:

– skewed condition: 8–2–2–2–2, which corresponds to a relative entropy of 0.86 / an entropy of 2;
– balanced condition: 4–4–4–2–2, which corresponds to a relative entropy of 0.97 / an entropy of 2.25.

[2] In the remainder of the chapter, I will continue to use *A* as referring to a word, but this should be understood merely as an attempt to avoid the hassle of highly repetitive disambiguation structures such as "the word/sense A" etc. – in fact, from now on *A* should be understood as referring to the locus of the most meaningful variation, which could be on the level of the word *or* the sense.

Probably somewhat surprisingly, they learned the new construction better from the skewed condition (and they found the same for adults in an earlier study). One way in which corpus linguists can quantify this easily is with the notion of (relative) entropy, which can be understood as a dispersion measure for categorical data or as a measure of uncertainty that comes with a distribution.

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i \text{ (with } n = \text{number of categories/types and } \log_2 0 = 0) \quad (3)$$

$$H_{relative} = H/\log_2 n \quad (4)$$

In this case, the skewed condition has a (relative) entropy of (0.86) 2 whereas the balanced condition has a (relative) entropy of (0.97) 2.25. In other words, the skewed condition exhibits a lower dispersion and less uncertainty, which makes it the one easier to learn, which is exactly what Goldberg (2006: 85f.) goes on to argue on the basis of additional and compatible findings from non-linguistic categorization that it is low-variance samples (recall that $H_{(relative)}$ is also a dispersion measure) that give rise to learning and categorization best. It is therefore not surprising that Zipfian distributions, in which a few types are highly frequent and, thus, reduce the entropy of the distribution, facilitate learning so much (see again Ellis and Ferreira-Junior 2009, in particular, 216).

Unfortunately, there is as yet very little work in corpus linguistics that takes distributions of token frequencies seriously enough. One interesting application is Mason's (1999) work on gravity, in which he uses entropy to identify which slots around a word exhibit how much variability. Another application closer to the current thread is Daudaravičius and Marcinkevičienė's (2004) work on the collocation measure of lexical gravity *G*. This measure does not only take the four token frequencies in Table 3.3 into consideration as nearly all other association measures, but also the number of types that goes into the *b* cell. On the one hand, this is approach still not sufficiently precise because for Table 3.4 it only adds the information that there are four contexts/types, but not their token frequencies let alone their entropies as in the above discussion of (3) and (4), which *G* would treat as the same although the entropies show they are not (see Gries 2012 for more detailed discussion). Nevertheless, it is an exciting step ahead and the first results are encouraging as when *G* outperforms *t* in terms of register discrimination (see Ferraresi and Gries 2011; Gries 2010a; Gries and Mukherjee 2010).

From all this, lesson 6 is as follows:

Lesson 6: Do not lump together contexts of (co-)occurrence but distinguish them and their type–token distributions and consider their dispersion/uncertainty.

Table 3.5 *Schematic representation of 'the frequencies of A in P (Q, R, S . . .) in X (Y, Z . . .)'*

	Context P	Context Q	Context R	Context S	Corpus part
A	66	23	16	7	X
A	35	33	15	8	Y

3.2.6 The frequency of A in P (Q, R, S . . .) in X (Y, Z . . .)

It is time to 'zoom out' some more. So far we have been concerned with increasing the resolution of our analytical procedures within a corpus (part) X. However, even the increased level of resolution advocated above comes with some risks, namely the risk that corpora are usually not very homogeneous entities. We have seen in Section 3.2.1 that, for instance, frequencies of A in a corpus X are a bold simplification because A's distribution may vary very much in files or in linguistically relevant corpus parts such as modes, registers, sub-registers, etc. On some level, corpus linguists are of course aware of this issue and sometimes report results for parts of corpora, especially for the above-mentioned pet distinction of speaking vs. writing, a scenario that can be schematically represented as in Table 3.5.

However, I have already shown above that picking any one level of resolution runs the risk that this is neither the quantitatively most revealing level (because it features the largest amount of variability to be explained) nor the linguistically most interesting level (because it supports the furthest-reaching linguistic generalizations). Section 3.2.1.2 therefore proposed to study multiple levels of resolution at the same time to be able to make an educated decision regarding which level of corpus organization to focus on. The present section presents a different approach, one that aims at suggesting to the analyst the best division of corpora in a bottom-up fashion by means of exploratory statistics such as cluster or principal components analyses.

One question one may be concerned with is how constructions' use differs across corpora or corpus parts, specifically, which verbs they are used with. As an example, consider the question how ditransitive constructions are used in different parts of the ICE-GB (from Gries 2011). Using the collostructional co-occurrence approach of Section 3.2.3, I computed for each verb that is used ditransitively at least once in the corpus (*give*, *show*, *send*, *allocate*, *award*, *cost*, and dozens of other verbs), how much it is attracted to, or repelled by, the ditransitive. Crucially, however, I did this for 18 different but overlapping corpus parts:

- the whole corpus (i.e., the traditional approach);
- the five registers: spoken dialog, spoken monolog, spoken mix broadcast, written printed, and written non-printed;

– the twelve sub-registers: private dialog, public dialog, scripted monolog, unscripted monolog, printed academic, printed creative, printed instructional, printed non-academic, printed persuasive, printed reportage, non-printed letters, non-printed non-professional.

These data were then arranged in a table with 18 columns (the corpus parts) and 87 rows (one for each verb used ditransitively), where the cells contain an association measure that reflected the mutual attraction/repulsion of the verb to the ditransitive in the relevant corpus part.

Data like this can be analyzed with a principal component analysis, an exploratory method that tries to, so to speak, compress the 18 columns into a smaller number of new columns (called principal components) by capitalizing on the similarities between columns. The theoretically extreme possible results are either that all 18 columns get compressed into one principal component because they are all so similar to each other, or that the 18 columns cannot be compressed at all and need to be retained because they are all so different from each other.

In this case, the analysis returned four orthogonal columns (which together account for nearly three-quarters of the variance in the data), which upon closer inspection revealed the following principal components:

– one component representing all spoken data but private dialog;
– one component representing spoken private dialog;
– one component representing written printed data;
– one component representing written non-printed data.

Thus, a corpus linguist interested in A's behavior in P, here how the ditransitive's verb slot is populated, should probably take register variation into consideration, but not necessarily the division of the corpus into (sub-) registers, but the above division into four components, because bottom-up analysis has shown that these are the components or corpus parts that, with regard to ditransitive verbs, behave most homogeneously internally and most heterogeneously externally. It is worth pointing out that this division of the ICE-GB is not a division along the lines of any of the divisions the corpus compilers made: this is not just spoken vs. written, or a division on the level of the registers or sub-registers. Rather, the bottom-up analysis shows that one needs to cut across all three levels of corpus analysis to arrive at the most accurate register account. Linguists do not usually like to mix different levels of categorization, but here their data clearly support exactly that. From this, lesson 7, a modification of lesson 2, follows:

Lesson 7: Corpus findings should not just be checked on multiple levels of corpus granularity but also explored with an eye to identifying the most discriminative, and thus potentially most interesting, division of the corpus.

3.2.7 The similarity of As in P (Q, R, S . . .) in X

In the discussion of how to do more justice to the complexity of frequency data, we have so far zoomed out: we started from A in X and successively extended the range of data to A in P (R, S . . .) in X (Y, Z . . .). At this point, we are 'zooming in' and look not just at the frequency of A in P in X, but also consider the individual matches and their characteristics. Why would one want to do this? Three reasons and/or areas of application come to mind: for example

- similarity is a driving force in first language acquisition: children do not hear something (multiple times) and immediately generalize to very different utterances – rather, their first non-repeated uses of a newly learned linguistic expression will in general be very similar to what they heard being used; see the Traceback approach pursued in, e.g. Dąbrowska and Lieven (2005), Lieven *et al.* (2009), Vogt and Lieven (2010);
- analogy and similarity are central to language change and grammaticalization (see Bybee 2010);
- syntactic priming, or persistence – i.e. speakers'/writers' tendency to reuse syntactic structures they have produced/comprehended not too long ago – is sensitive to similarity between utterances.

It is two examples for priming/persistence that I want to discuss here because they give rise to a somewhat scary implication. First, a set of examples of what I will call *local similarity*. Szmrecsanyi (2005, 2006) explores two kinds of persistence:

- α-persistence: a structure *x* increases the probability of the same structure *x* at the next point where *x* competes with a functionally similar structure;
- β-persistence: a structure *x* increases the probability of a similar structure *y* at the next point where *x* competes with a functionally similar structure.

It is the latter kind of persistence that is of interest here because – unlike the former – it does not involve identity of structures, but just their similarity. Of the range of case studies Szmrecsanyi discusses, I will single out three, one on analytic vs. synthetic comparisons, one on *will-* vs. *going-to* futures (both based on data from the *BNC*), and one on particle placement (i.e., the alternation between *I threw away his iPhone* and *I threw his iPhone away*, based on data from the *Freiburg Corpus of English Dialects*). In all cases he finds significant effects of β-persistence/similarity: comparison *more* (e.g., *I like Linux more than Mac*) primes – i.e., increases the probability of – analytic comparatives with *more*; *go* in motion senses (e.g., *I would never go into a Mac store*) primes *going-to* futures; when the same phrasal verb is used, particle placement is primed more strongly (see also Gries 2005). Thus, even in the

face of syntactic differences, lexical similarity results in, or more conservatively facilitates, reactivation and reuse of syntactic structures.

While these are very interesting findings, one problem remains: "the problem ... is in specifying the relevant features upon which similarity is measured" (Bybee 2010: 62). One approach to this question is to not just use individual features but adopt a view of similarity I will refer to as *global similarity*. One case study in Snider (2009) does just that. He tested whether the overall similarity between a prime and a later point of choice, the target, facilitates priming. Using examples of the dative alternation in the Switchboard corpus that were annotated for a large number of characteristics that influence the choice of a ditransitive or a prepositional dative, he computed the pairwise similarities of primes and targets with the Gower's multi-feature distance metric shown in (5) (which is implemented in R as the function daisy, in the library cluster):

$$d_{ij} = \sum_{k=1}^{p} w_k \delta(ij;k)\, d(ij;k) \div \sum_{k=1}^{p} w_k \delta(ij;k) \qquad (5)$$

In a nutshell, this metric compares categorical variables for identity and numeric variables for normalized differences. More importantly, however, is that he found that the more similar prime and target are to each other (all other things considered), the more likely they are to also involve the same construction. This can be schematically represented as in Table 3.6, where matches of *give* in ditransitives and prepositional datives are listed aligned according to slots whose similarities may be compared.

There are two conclusions to draw from this. One is fairly straightforward: not only does similarity play a role, it does so on very many different levels of analysis. The other follows from this but is more scary: the key to understanding what happens in a particular concordance line at time *t* may involve (much) less of what is happening in that particular instance at time *t* rather than what happened before *t*, namely

- at the last point(s) before *t* the speaker/writer had to make a choice between the same kinds of constructions (α-persistence);
- any time before *t* when something was produced/comprehended, that is, on any number of levels, similar to the choice the speaker has to make now at *t* (β-persistence)!

Table 3.6 *Schematic representation of 'the similarity of As in P (Q, R, S ...) in X'*

match 1 (in X)	He		**gave**	him	the book	
match 2 (in X)	My father	did not	**give**		the car	to him
match 3 (in X)			**Give**	peace	a chance	
match 4 (in X)	The mailman		**gave**	the guy	the finger	

That is to say, the choice of an analytic comparison in a line of one's concordance may not make any sense at all given everything that one can see in that utterance with the analytic comparative – the adjective may be frequent, have final stress, and be used attributively, which are all features promoting a synthetic comparative – and may only make sense if one recognizes that, five clauses earlier, the speaker used *more* ... Plus, not only will both of these persistence effects be amplified or reduced by the degree of similarity between what happened before, both of these effects are potentially long-lasting. Both Gries (2005) and Szmrecsanyi (2005, 2006) found priming effects across many intervening clauses (priming decays logarithmically over time), and similar findings have been made in experimental designs (see Bock and Griffin 2000, who found priming across ten sentences). Thus, although corpus linguists already often have to scrutinize many matches/concordance lines, the somewhat frustrating last lesson is the following:

Lesson 8: To understand/explain speakers'/writers' choices at one point of time, it is often indispensable to explore quite some preceding context for influences from previous lexically/structurally identically choices as well as 'only' similar constructions.

This point has even more important implications. This also means that it is risky to generate a concordance but then only sample, say, 10 percent of this concordance for more detailed annotation and analysis. This is because unless a large context is included for every concordance match, the individual matches will be divorced from their context, which, as we have just seen, may contain the only good explanation of what a speaker is doing. Thus, any such sampling from a concordance requires that the analyst include a large amount of context for analysis.

3.3 Concluding remarks

I hope to have shown how risky an overly simplistic view and treatment of frequencies can be. Many studies in cognitive/usage-based linguistics have shown that speakers keep track of vast amounts of multidimensional and probabilistic co-occurrence information, and by now it is also well understood how early this begins – in fact, such learning processes begin *in utero* – and how fast this happens – speakers can pick up meaningless but probabilistically somewhat reliable patterns after just a few minutes of input. It is therefore only prudent to have our work with frequency data be similarly comprehensive – rather than just focusing overly narrow on simple frequencies of occurrence or co-occurrence in a corpus in general or in a narrowly defined context, we must be more aware of all the factors that pose serious threats to our analyses and that we know that speakers, whose behavior we ultimately want to explain or at least describe, respond to. Obviously, for

many real-life constraints, we will not be able to always consider all of the above aspects of (co-)occurrence frequencies, but we must understand that that is what is theoretically necessary.

While I have not been concerned much with different kinds of statistics in this chapter, I hope this chapter also makes it obvious why multifactorial analyses are the future. Even if one is interested in only the most basic frequency data of the least complex phenomena (whatever linguistic phenomenon *P* would ever be simple is not clear to me, but let's assume for the moment, there was one), the above argumentation should have shown that one probably needs to consider

- *P*'s dispersion across files;
- *P*'s dispersion across registers, for which a multivariate bottom-up exploration can be useful;
- *P*'s alternative realizations or contexts, for which *n*-dimensional tables may have to be considered;
- *P*'s larger context, which means similarities and priming effects have to be measured and considered.

That is a lot of work . . . but it's in fact good news because it shows that corpus linguistics is coming of age in terms of now having a range of methods that can shed light on things in a way that does justice to the complexity of the speakers' linguistic systems generating the data, and in terms of allowing us to connect to neighboring disciplines that have been using corpora but often not as well as they could have, cognitive linguistics being a case in point. On that note, I hope this chapter will help in pursuing our research goals with renewed methodological rigor and interdisciplinarity.