Surdeanu, M., R. Morante and L. Màrquez (2008): "Analysis of joint inference strategies for the semantic role labelling of Spanish and Catalan," in *Lecture Notes in Computer Science*, 4919, 206-218.

Taulé, M., M. Martí and M. Recasens (2008): "AnCora: multilevel annotated corpora for Catalan and Spanish," in *Proceedings of LREC 2008*.

Vaamonde, G. (2008): "Algunos problemas concretos en la anotación de papeles semánticos. Breve estudio comparativo a partir de los datos de AnCora, SenSem y ADESSE.," *Procesamiento del lenguaje natural*, 41, 233-240.

Useful statistics for corpus linguistics

Stefan Th. Gries
University of California, Santa Barbara

1. Introduction

By its very nature, corpus linguistics is a distributional discipline. In fact, it has been argued that corpora as such contain nothing but distributional frequency data, more specifically distributional data of two or three types, depending on how one wants to look at it:

a) frequencies of occurrence of linguistic elements, which can be studied from two different perspectives:
    - how frequent are morphemes or words or patterns/constructions in (parts of) a corpus? This information can be provided in various different forms of frequency lists;
    - how evenly are morphemes or words or patterns/constructions distributed across (parts of) a corpus? This information can be provided in the form of various dispersion statistics;
b) frequencies of co-occurrence: how often do linguistic elements such as morphemes, words, patterns/construction co-occur with another linguistic element from this set or a position in a text.

In other words, in the first instance and from a purist's perspective, all a corpus returns is whole numbers greater than or equal to 0, namely frequencies of how often something occurred in a corpus. Everything else that (corpus) linguists are actually interested in must then be operationalized on the basis of some kind of frequencies.

While the degree of commitment to this a bit more radical view may be subject to debate, it is nevertheless probably fair to say that, for these reasons, branches of linguistics that have been using corpora or text databases have always been among the most quantitatively oriented subdisciplines of the field. While this would typically lead to the expectation that corpus linguists are very much statistically-minded – after all, statistics is the scientific discipline that teaches us how to deal with quantitative distributions – it is unfortunately also fair to say that not all corpus-based studies utilize the available statistical methods to their fullest extent. In fact, it is only in the last few years or so that corpus linguists have begun to use more, more sophisticated, and more comprehensive tools, both for handling of corpus data as well as for the statistical analysis of the

resulting data. However, this trend is new enough for the discipline to not have evolved to a state where such resources are widely used and taught, and so far there is only one dedicated introduction to statistics for corpus linguists (Oakes 1998), which in spite of its merits at the time of publication is beginning to be a bit dated, plus one introduction to corpus linguistics with a detailed overview chapter on statistical methods (Gries 2009a).

In this chapter, I provide a brief survey of several statistical concepts and methods that are relevant to two stages of corpus-linguistic applications. Sections 2 discusses basic concepts regarding the kinds of distributional data mentioned above as well as very simple descriptive methods to use such data: absolute and relative frequencies of occurrence (Section 2.1), the distribution of linguistic elements across a corpus (Section 2.2), and statistical measures for frequencies of co-occurrence (Section 2.3). More specifically, I discuss several widely-used measures or methods, but then also point to additional less widely used methods which I think the field as a whole should consider more often.

Section 3 deals with statistical methods whose purpose is to evaluate the data obtained from the methods in Section 2 and which often involve significance levels and *p*-values. Once frequencies of (co-)occurrence have been obtained, the corpus linguist typically wants to explore and describe the interrelation(s) between what may be causes and effects in a monofactorial or a multifactorial way (Sections 3.1 and 3.2 respectively) or detect structure in large and often messy corpus data (Section 3.3).

Limitations of space require two comments before we begin. First, I cannot discuss the very foundations of statistical thinking, the notions of (null and alternative) hypotheses, (independent and dependent) variables, Occam's razor, etc., but must assume the reader is already familiar with these (or reads up on them in, say, Gries 2009a: Ch. 5). Second, most of the time I cannot exemplify the relevant methods in detail. The number of collocational statistics (25+), the number of dispersion measures and adjusted frequencies (20+), the technicalities of logistic regression or cluster analysis make it impossible to, for example, provide practical examples with the statistical code to analyze that example, which is why I will provide references for further reading in each (sub)section.

2. Distributional data from corpora
2.1. *Frequencies of occurrence*
The most basic corpus-based statistic is of course the *observed absolute frequency* of some phenomenon. For example, looking for the word forms *give* and *bring* in the British Component of the International Corpus of English returns 441 and 197 matches respectively. A similarly simple example involves the observed frequencies within a part of the corpus: *give* occurs 297 and 144 times in the spoken part and the written part respectively whereas *bring* occurs 128 and 69 times in the spoken part and the written part respectively. Obviously, there

are more occurrences of *give* in this corpus than *bring*; obviously there are more occurrences of *give* and *bring* in the spoken component of the corpus than in the written component of the corpus

2.1.1. Normalization, comparison, and logged frequencies
While this is so basic as to hardly merit discussion, it is a point of entry for some other concepts. First, it is necessary to point out that a higher frequency occurrence of some element in some corpus (part) does not automatically that the element observed more often is more frequent because the observed frequencies are of course dependent on the sizes of the corpus parts that are compared. In this case, it is possible to directly compare the observed absolute frequencies of *give* and *bring* in the ICE-GB, but it is not possible to directly compare the observed absolute frequencies of *give* and *bring* across the spoken and written components because these two components are not equally large. Instead, what is needed are the *observed relative frequencies*, which are typically normalized and reported as frequencies per 1,000 or 1,000,000 words.

In this particular case, the ICE-GB contains 1,061,263 words, which means that the observed relative frequencies of *give* and *bring* in the whole corpus become those in (1), and for the above reason the ratio of these two relative frequencies is identical to those of the absolute frequencies:

(1) a. *give* (whole corpus): $\dfrac{441 \cdot 1000000}{1061263} \approx 415.54$

  b. *bring* (whole corpus): $\dfrac{197 \cdot 1000000}{1061263} \approx 185.63$

However, the situation changes for the spoken and the written data. The spoken and the written components contain 637,682 and 423,581 words respectively. Thus, the relative frequencies for *give* and *bring* in the spoken and written data become those shown in (2), and it is obvious that just because the observed absolute frequency of *give* in speaking is about two times as high as that in writing does not mean that this is the actual ratio between *give*'s frequencies in speaking and in writing, which is approximately 1.37:

(2) a. *give* spoken: $\dfrac{297 \cdot 1000000}{637682} \approx 465.75$, and written: $\dfrac{144 \cdot 1000000}{423581} \approx 339.96$

  b. *bring* spoken: $\dfrac{128 \cdot 1000000}{637682} \approx 200.73$, and written: $\dfrac{69 \cdot 1000000}{423581} \approx 162.9$

It is therefore important to bear in mind that one can only compare corpus frequencies or use them to make statements about what is more frequent when the frequencies have been normalized.

Second, relative frequencies can be used to compare different corpora with each other just by computing the *relative frequency ratio*, the quotient of the relative frequencies of a word in both corpora (Damerau 1993). For example, consider the Wikipedia entries on the two multi-purpose programming languages Perl and Python, which at some point of time contained 6,065 and 5,596 words respectively. Table 1 shows the 10 largest and smallest relative frequency ratios for words occurring in both entries (with their frequencies of occurrence). A statistically more elaborate way to compare corpus frequencies, which often yields similar results, has been popularized as the method of key words (cf. Scott 1997).

Words characteristic for Perl			Words characteristic for Python		
Word	Frequencies	RelFreqRatio	Word	Frequencies	RelFreqRatio
perl	249 : 8	28.72	*python*	3 : 208	0.01
source	14 : 1	12.92	*classes*	1 : 12	0.08
interpreter	27 : 2	12.46	*set*	1 : 11	0.08
statement	13 : 1	11.99	*objects*	1 : 9	0.09
regular	24 : 2	11.07	*within*	1 : 8	0.1
returns	11 : 1	10.15	*function*	2 : 14	0.12
structures	10 : 1	9.23	*numbers*	1 : 7	0.13
write	9 : 1	8.3	*well*	1 : 7	0.13
tasks	9 : 1	8.3	*style*	2 : 13	0.13
community	9 : 1	8.3	*dictionary*	1 : 6	0.14

Table 1. Highest and lowest relative frequency ratios for the Wikipedia entries for *Perl* and *Python*

Third, it is also worth mentioning that there are also a variety of contexts in which one does not use the observed absolute or relative frequencies of linguistic elements but the logarithms of their observed absolute frequencies. This is especially useful when corpus frequencies are correlated with other data such as data from psycholinguistic experiments (e.g. reaction times). The log transformation – usually to the base of 2, sometimes to the bases of *e* (2.71828) or 10 – has the effect that the otherwise very skewed distribution of word frequencies is linearized and allows us to use simpler linear correlation measures to compare the corpus frequencies with other data.

2.1.2. Further hints: useful but underutilized statistics
One very useful statistic is *entropy H* or *relative entropy* H_{rel}. Entropy is the average amount of uncertainty of a random variable: the larger H or H_{rel}, the more random a distribution is and, at the same time, the more difficult it is to predict an element's occurrence. The entropy and the relative entropy of a distribution with *n* observations are defined as in (3):

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

     b. $$H_{rel} = H / H_{max} = H / \log_2 n$$

Since H is correlated with *n*, H_{rel} allows to compare entropies of distributions of different *n*s. For example, the frequencies of the inflectional forms of the two verbs *give* and *sing* in the ICE-GB are as follows:

give: 441	*gives*: 105	*giving*: 132	*gave*: 175	*given*: 376
sing: 38	*sings*: 2	*singing*: 45	*sang*: 3	*sung*: 2

Table 2. Observed frequencies of *give*'s and *sing*'s inflectional forms in the ICE-GB

If one knew the above frequencies of *give* and was asked to predict what the next form of the lemma *give* in a corpus would be, one would guess *give*, because that form is the most frequent form. However, there would be a relatively large amount of uncertainty because *given* is also fairly frequent, and even the other verb forms are not at all infrequent. With the lemma *sing*, the situation is different: one would guess *singing*, but in general the uncertainty is lower: while *sing* is also relatively frequent, the other three forms are very infrequent. The entropy values capture that and show that the average uncertainty for the lemma *give* is larger: H for the lemma *give* is 2.1 (H_{rel}=0.91) while H for the lemma *sing* is only 1.4 (H_{rel}=0.62). Entropies are an interesting and revealing way to summarize the randomness of distributions that has many applications (one of which will be discussed below).

One problem that arises often especially when one uses smaller corpora is that of observed frequencies of zero, i.e. when a particular linguistic expression is not observed at all. While the non-occurrence of an expression may in some cases be meaningful, in fact even expected, in other cases it is not meaningful and/or can be problematic for statistical estimation purposes. A rather ingenuous method to estimate the probability of unseen items, and accordingly adjust the probability of all attested items, is called Good-Turing estimation. While the underlying mathematics are somewhat complicated, Gale and Sampson (1995) provide an accessible introduction to this approach (as well as R code for its implementation).

Another final, more frequent and potentially much more threatening problem is based on the fact that, strictly speaking, observed frequencies and all statistics based on them can in fact be very misleading. This threat is the topic of the next section.

2.2. Dispersion and adjusted frequencies

For what follows, it is useful to briefly realize what an observed relative frequency is. Imagine a corpus consisting of three parts of seven elements each that is shown in (4), where the corpus parts are delimited by "|":

(4) q w w e e e r | q r r t t t t | q y y y y y y

The observed absolute frequency of y in this corpus is 6, the observed relative frequency is $^6/_{21}$=28.57%. A moment's thought will reveal that this relative frequency is in fact a mean, namely the mean one of the distribution when every y is changed to 1 and every other character to 0:

(5) 0 0 0 0 0 0 0 | 0 0 0 0 0 0 0 | 0 1 1 1 1 1 1

As one learns in every introduction to statistics, "never provide a mean without also providing an index of dispersion." This rule applies here, too. The overall mean does not reveal that the distribution of y across the three is extremely uneven: While y's overall relative frequency is 28.57%, its relative frequencies across the corpus varies between 0 and 85.71%, and neither of these frequencies is summarized well by the overall mean. Note how q, on the other hand, has a perfectly regular distribution: its overall relative frequency is 14.29%, as is its relative frequency in each corpus part.

Situations like these can not only be rather frequent but can also potentially undermine any statistic based on frequencies. If an overall corpus frequency is high, then this will be reflected in whatever other statistic is based on that frequency, but if the high frequency is completely unrepresentative of most of the corpus, the results will be of dubitable value only.

In order to cope with this kind of problem, two different but related solutions have been proposed: on the one hand, one can quantify the degree of *dispersion* of the counted linguistic expression in the corpus so that one knows how well the corpus frequency reflects the expressions overall distribution. On the other hand, one can downgrade the observed frequency to a so-called *adjusted frequency* in proportion to the degree that the expression in point is unevenly distributed in the corpus.

There is too large a number of dispersion measures and adjusted frequencies here to discuss them all, and unfortunately there is also as yet little work, let alone agreement, on which measure is best (cf. Gries 2008, 2009c), which is why I will just mention the most recently proposed dispersion measure, DP_{norm} (for normalized deviation of proportions). This measure ranges from 0 to 1, where values around zero mean that the relative frequencies of occurence of the linguistic expression in the corpus are directly proportional to the sizes of the corpus parts whereas large values close to 1 mean the linguistic expression is

distributed very unevenly among the corpus parts. For y and q, DP_{norm} amounts to 1 and 0 respectively: y's maximally uneven distribution is clearly reflected.

The importance of (under)dispersion cannot be overestimated. If, as mentioned above, every corpus statistic ultimately relies on frequencies, then frequencies based on uneven distributions can undermine all subsequent analytical steps, which is why measures of dispersion should be provided for virtually all corpus frequencies.

2.3. Frequencies of co-occurrence

The second most basic corpus statistic involves frequency of co-occurrence of linguistic expressions. Unfortunately, the existing terminology is not used consistently by different researchers; the following are kinds of co-occurrence that are distinguished:

- co-occurrence of words (*collocation*);
- co-occurrence of words and constructions/patterns (*collostruction* or one sense of *colligation*);
- co-occurrence of constructions/patterns (another sense of *colligation*);
- words and textual positions (another sense of *colligation*).

The most basic way of providing co-occurrence information is again just providing the observed absolute frequency. For example, in the ICE-GB *give* and *bring* (words) occur as a ditransitive verb (the pattern) 232 and 2 times respectively. However, since we already know that *give* and *bring* are differently frequent in this corpus (cf. above), the absolute observed frequencies are not the meaningful statistic to report, and the next section will explore more suitable alternatives.

2.3.1. Expected frequencies and association measures

Since the absolute frequencies cannot be directly compared, what is needed is a measure that takes the overall frequencies of *give* and *bring* into consideration. The most frequent strategy is to instead report a statistical *association measure* between the words and the construction. More than 20 such measures have been proposed in the literature (cf. Evert 2004 and Wiechmann 2008), but nearly all of them are based on 2×2 co-occurrence tables of the kind exemplified in Table 2 for *give*. For such a table, one needs the frequency of *give* (441), the frequency of *give* as a ditransitive verb (232), the frequency of ditransitive verbs (1,803), and the frequency of all verbs (138510), which are italicized in Table 3, from which the missing frequencies can be computed by subtraction.

	Construction ditransitive	Not construction ditransitive	Totals
Word *give*	232	209	441
Not word *give*	1571	136498	138069
Totals	1803	136707	138510

Table 3. Observed co-occurrence frequencies of *give*
and the ditransitive in the ICE-GB

The most frequent measure more useful than the observed frequency of 232 include *pointwise Mutual Information*, the *t*-score, and maybe -$\log_{10} p_{\text{Fisher-Yates exact}}$ $_{\text{test}}$. They all require to first compute the frequencies one would expect if give and the ditransitive occurred together only randomly, which are computed as shown in Table 4.

	Construction ditransitive	Not construction ditransitive	Totals
Word *give*	$\dfrac{441 \cdot 1803}{138510} \approx 5.7$	$\dfrac{441 \cdot 136707}{138510} \approx 435.3$	441
Not word *give*	$\dfrac{138069 \cdot 1803}{138510} \approx 1797.3$	$\dfrac{138069 \cdot 136707}{138510} \approx 136271.7$	138069
Totals	1803	136707	138510

Table 4. Expected co-occurrence frequencies of *give*
and the ditransitive in the ICE-GB

On the basis of these frequencies, the three association measures are:[1]

(5) pointwise *MI*: $\log_2 \dfrac{\text{observed frequency}}{\text{expected frequency}} = \log_2 \dfrac{232}{5.7} \approx 5.3$

(6) *t*: $\dfrac{\text{observed frequency - expected frequency}}{\sqrt{\text{observed frequency}}} = \dfrac{232-5.7}{\sqrt{232}} \approx 14.9$

(7) -$\log_{10} p_{\text{Fisher-Yates exact test}}$: $-\log_{10}(8.7 \cdot 10^{-315}) \approx 314.1$

If pointwise *MI* and *t* are positive, then the observed frequency of co-occurrence is larger than the expected by chance, and the larger all association measures, the stronger this effect is. Without much experience, these values are hard to interpret, but these are all rather high values which point to a very strong posi-

---

1   The formula for the Fisher-Yates exact test involves summing multiple *p*-values derived from the hypergeometric distribution and is too complex to be provided here; cf. the above references.

tive association between *give* on the one hand and ditransitives on the other hand; this is probably easiest to see from the very large difference between the observed frequency of 232 and the very small frequency expected by chance of 5.7. Usually, however, analogous computations are performed for (many) different verbs, and the values for *give* are set into perspective by comparing them to those for *bring*:

(8) pointwise *MI*: $\log_2 \dfrac{\text{observed frequency}}{\text{expected frequency}} \approx -0.4$

(9) *t*: $\dfrac{\text{observed frequency - expected frequency}}{\sqrt{\text{observed frequency}}} \approx -0.4$

(10) -$\log_{10} p_{\text{Fisher-Yates exact test}}$: $-\log_{10}(0.53) \approx 0.28$

The negative values for pointwise *MI* and *t* in (8) to (10) show that *bring* occurs less often in the ditransitive (2) than would be expected by chance (2.6), but all measures show that this dispreference of *bring* to occur in the ditransitive is not strong.

The most frequent application of such co-occurrence measures is in fact not in the domain of collostructions/colligations, but in the domain of collocations, i.e., the association between words. The overall logic is the same: the co-occurrence frequency becomes the co-occurrence frequency of the two words, the overall frequencies of the word and the construction become the overall frequencies of the two words, and the frequency of all verbs becomes the corpus size in words; the association measure is then interpreted the same way.

2.3.2. Further hints: properties of association measures

Just like with the dispersion measures, a large number of association measures has been proposed. Unlike with dispersion measures, however, the properties of association measures are better known. In general, it is probably fair to say that association measures often behave similarly, but unfortunately they can still output very different results and it is therefore necessary to bear a few things in mind when choosing a measure.

First, it is important to recognize that association measures are in general sensitive when it comes to very low frequencies. In such cases, association measures will often overestimate the association between elements even though the small frequencies of the elements would not support a high degree of confidence into a strong association.

Second, some association measures are asymptotic in nature and, thus, presuppose a particular distribution of the data (at least when they are used for significance testing), but, for example, corpus data are usually not normally distributed. One must therefore be careful to properly justify the choice of a

measure or, maybe safer, choose measures that do not make such distributional assumption (such as the Fisher-Yates exact test or the *binomial test*).

Third, given the different mathematical properties of the association measures, it is important to know that they are differently sensitive to low frequencies and will also rank-order collocations differently depending on the words' frequencies. For example, pointwise *MI* is known to return very high association scores to low-frequency words as well as technical terms or other expressions that exhibit very little or no variation. On the other hand, the *t*-score returns high association scores to word pairs with high co-occurrence frequencies and provides a better measure of the non-randomness of the co-occurrence (cf. Evert 2009b for more discussion).

2.3.2. Further hints: some desiderata and underutilized approaches
Let us complete this section with a few desiderata.

First, nearly all of the association measures are based on token frequencies only. That is, the frequencies entered into Table 2 do not reveal anything about the number of different verbs that make up the 1571 non-*give* ditransitives. On the one hand, this makes the association measures much easier to compute, but on the other hand, they lose information. Obviously when *give* occurs in the ditransitive 232 times and the remaining 1751 ditransitives are made up of 300 other verbs, then this speaks more in favor of a strong association of *give* to the ditransitive than when the remaining 1751 ditransitives are made up of 3 other verbs. Unfortunately, I know of only one association measure that incorporates such information – Daudaravičius and Marcinkevičienė's (2004) gravity counts – and in spite of its uniqueness and promise, this measure has so far hardly been explored (but cf. Mukherjee and Gries 2009).

Second, while the study of collostructions, colligations, or collocational frameworks (i.e., patterns such as *a ___ of*) involves the co-occurrence of a word with a construction in a precisely defined syntactic slot, many approaches to collocation rather adopt a window-based approach in which all words in a window of, often, four or five words around the relevant node word are considered. Again, this is computationally easy, but again it loses information. A better but also underutilized approach is Mason's (1997, 1999) lexical gravity. (Daudaravičius and Marcinkevičienė's (2004) gravity counts are based on this work by Mason). Mason proposes to explore a much larger window of slots around a node word by computing, for each slot around a word, the entropies of the collocates in that slot to determine which positions around a node word of interest exhibit how much variation. This approach is an interesting way to determine the relevant contextual slots of words in a bottom-up fashion and readily allows to consider asymmetrical windows around node words.

Another interesting set of problems arises when the notion of collocation is applied to collocations of more than two words. (When *n* words in question form

a chain, they are often referred to as *n*-grams). One problem is that many of the statistical measures that have been proposed for the collocation of two words cannot straightforwardly be applied to collocations of more than two words or *n*-grams with *n*>2. For instance, one problem involves the question of how to compute the expected frequencies: should the expected frequency of *in spite of* be based on independent probabilities of *in*, *spite*, and *of*, or on the independent probabilities of *in spite* and *of*, or on the independent probabilities of *in* and *spite of*? Commercially available software usually computes expected frequencies in the first of these three ways, but it is not clear (yet) whether this is really the appropriate way.

Another often underestimated problem is even more fundamental, namely the question of which *n* to assume for *n*-grams. For example, it is obvious that it would rarely be useful to treat *in spite* as a 2-gram – the relevant *n*-gram is the 3-gram *in spite of*. For example, it may be less useful to treat *the one hand* as a 3-gram, because the 'real' *n*-gram may well be *on the one hand*, etc. Most studies using *n*-grams have so far defined an *n* a priori but it may often be more prudent to let the corpus data decide in a bottom-up, *n*-gram specific fashion what the best statistically best *n*-gram sequences are.

A final corpus-linguistically completely understudied problem related to the previous one involves discontinuous *n*-grams, the identification of which is not only statistically difficult in the above way, but also computionally extremely demanding. It remains to be hoped that corpus linguists will address these topics (more) soon and consider work already undertaken in computational linguistics (cf. Kita et al. 1993, 1994; Nagao and Mori 1994; Ikehara, Shirai, and Uchino 1996; Shimohata, Sugio, Nagata 1997; da Silva, Dias, Guilloré, and Pereira Lopes 1999; Mukherjee and Gries 2009).

3. Statistical evaluation of distributional data
In this section, I will discuss a few central statistical methods for the analysis of corpus data.

3.1. Correlations of interval/ordinal-scaled data
One frequent scenario involves the comparison of how two variables are related to each other, and such relations are sometimes, and should always, be explored graphically in a scatter plot of the kind exemplified in Figure 1, which uses a solid line to show the correlation between time (on the *x*-axis) and the relative frequency of the expression *just because* (on the *y*-axis) (from Hilpert and Gries 2009; we disregard the dashed line for now).
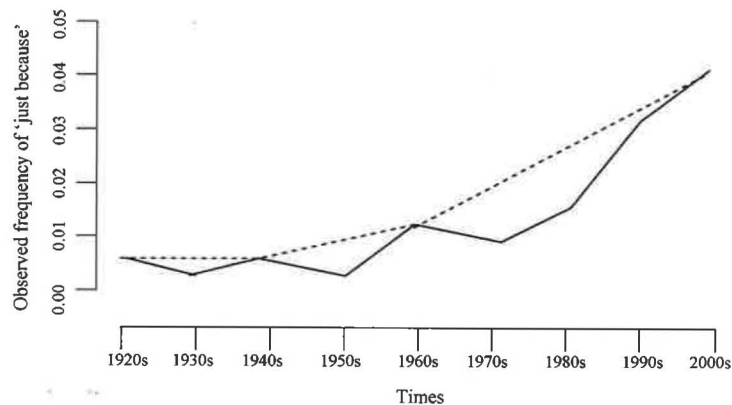
Figure 1. The development of the frequency of *just because* over time

It is obvious that there seems to be a positive trend, but of course this needs to be tested. The probably most frequently used correlation is Pearson's *r*, but the significance test for Pearson's *r* relies on distributional assumptions that corpus data often violate, so it is nearly always safer to use an alternative measure that does not, such as Kendall's *tau*. Like many correlation coefficients, *tau* is close to 1 when there is a strong positive correlation ('the more *a*, the more *b*'), it is close to -1 when there is a strong negative correlation ('the more *a*, the less *b*'), and it is close to zero in the absence of a correlation. Kendall's *tau* for the correlation between the time and the relative frequency is rather high and positive (τ=0.743; *z*=2.74; *p*=0.006), which means that, obviously, the frequency of *just because* increases significantly over time.

While Kendall's *tau* is often safer to use than Pearson's *r*, it can only be applied to scenarios in which two variables are correlated – for scenarios where more than one variable is involved in the prediction of a dependent variable, frequently the approach of *linear modeling* is used, which uses the same logic underlying Pearson's *r*. The mathematically simplest approach of linear models requires that all variables involved are interval-scaled, and the model selection process is typically performed in such a way that one begins to model the data with the most comprehensive model that one's data allow one to formulate, and then non-significant predictors are removed in a stepwise fashion (following Occam's razor), until only significant predictors remain. While corpus data often violate the assumptions of linear models, the logic of linear modeling and the model selection process helps to understand other statistical regression methods that can handle corpus data (cf. Section 3.2.2 below).

3.2. (Cross-)Tabulation of frequency data
3.2.1. Chi-square statistics

The simplest way of analyzing frequency data involves creating a one-dimensional frequency table of linguistic expressions and perform a test of whether the distribution of the observed frequencies of occurrence deviate significantly from an expected distribution, where that expected distribution may be just a random distribution or a distribution following from some other study or expectation.

For example, Table 2 above listed the frequencies of the inflectional forms of the lemma *give*, for which then also H and H_{rel} were computed. A *chi-square test* can now also show whether the observed frequencies differ significantly from a uniform distribution – i.e., the case where all inflectional forms are equally frequent. A chi-square test shows that the frequencies of the inflectional forms of *give* are not compatible with the assumption that they are equally frequent: χ^2=377.72; *df*=4; *p*<0.001.

Also, on the basis of a previous study one may believe that the frequencies of the five inflectional forms should be distributed as follows: 40% vs. 10% vs. 10% vs. 15% vs. 25%. Again, a chi-square test can show whether the observed frequencies fit that distribution, and they do not: χ^2=24.35; *df*=4; *p*<0.001.

The probably more frequent scenario is that one cross-tabulates two different variables with each other, to obtain tables of the kind exemplified in Table 3. For example, Hundt and Smith (2009) discuss the frequencies of the simple past and the present perfect in four corpora on the basis of the following data:

	LOB	FLOB	BROWN	FROWN	Totals
pres. perf.	4196	4073	3538	3499	15306
simple past	35821	35276	37223	36250	144570
Totals	40017	39349	40761	39749	159876

Table 5. Observed co-occurrence frequencies of two tenses in four corpora

Hundt and Smith (2009: 51) conclude that they are "dealing with stable regional variation rather than ongoing diachronic change." However, there are two potential problems. First, their way of representing the data is statistically speaking less than ideal; second, they do not provide a statistical test to support their interpretation. As for the second problem, a chi-square test shows that they are right: there is a significant correlation between the tenses and the corpus parts (χ^2=130.8; *df*=3; *p*<0.001), but the results also show that the two British and the two American varieties pattern together: BrE has more present perfects than expected, and AmE has less. This pattern can be very clearly seen in an *association plot* such as Figure 2, in which black and white bars indicate observed frequencies that are larger and smaller than expected by chance respectively.
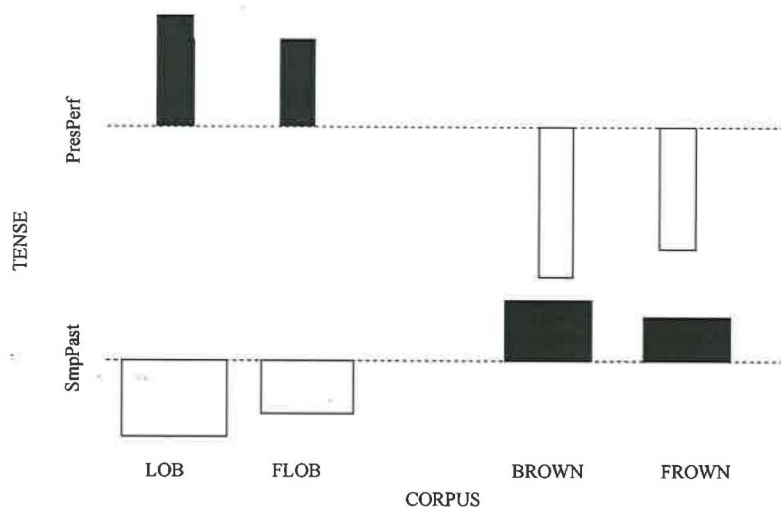
Figure 2. Association plot for the interaction TENSE:CORPUS

It should be noted, however, that the observed effect is also very small, as a standard *effect size* for *r*×*c* tables shows: Cramer's *V*=0.03. The highly significant result arises because of the large sample size.[2]

However, while the above computation exemplifies the use of the chi-square test (with the caveat of n. 2), there is still the first problem, which is somewhat more tricky and leads to the next section on multifactorial models. This problem is that one would normally actually not compute the above chi-square test on these data. Hundt and Smith represented the data as if it represents a two-dimensional data set: TENSE (present perfect vs. simple past) × CORPUS (LOB vs. Brown vs. FLOB vs. Frown) whereas in fact it is a three-dimensional data set: TENSE (present perfect vs. simple past) vs. VARIETY (BrE vs. AmE) vs. TIME (early (for LOB and Brown) vs. late (for FLOB and Frown)). Just because such data *can* be represented two-dimensionally does not mean they *are* two-dimensional. However, for such a three-dimensional data set, a chi-square test is not the most appropriate method – rather, a *generalized linear regression* is more appropriate here.

---

[2] Note that the above kind of chi-square test assumes that all data points are independent of each other, a condition which is strictly speaking probably not met because all the data points provided by one and the same author are related to each other; cf. Baroni and Evert (2009) and Evert (2006, 2009a) for insightful discussion of this problem as well as suggestions for analysis (cf. also Gries 2006a, b).

3.2.2. Generalized linear models: logistic regression and Poisson regression
Just like the linear models mentioned above, generalized linear models try to predict the outcome of a dependent variable on the basis of one or more independent variables. Also, they involve the same kind of model selection process in which non-significant variables are successively eliminated. The main difference to simple linear models is that the dependent variable need not be a interval-scaled variable but can also be factors with two levels (*binary logistic regression*) or more levels (*multinomial logistic regression*) or frequencies (*Poisson regression*).

To at least briefly exemplify one application this approach, we can revisit Hundt and Smith's (2009) data from the previous section and, first, represent them in a way that makes the design of the study more obvious. In Table 6, the left three columns contain what may be considered the independent variables while the rightmost column contains the observed frequencies of the levels of the independent variables.

TENSE	VARIETY	TIME	Frequency
present perfect	BrE	early	4196
present perfect	BrE	late	4073
present perfect	AmE	early	3538
present perfect	AmE	late	3499
simple past	BrE	early	35821
simple past	BrE	late	35276
simple past	AmE	early	37223
simple past	AmE	late	36250

Table 6. Observed co-occurrence frequencies of two tenses in four corpora

The model selection process of, in this case, a Poisson regression begins with a completely saturated model, which contains the three variables, their three two-way interactions (TENSE:VARIETY, TENSE:TIME, and VARIETY:TIME), and their three-way interaction (TENSE:VARIETY:TIME). However, this first model shows that the three-way interaction is not needed, and subsequent elimination of interactions shows that TENSE:TIME and VARIETY:TIME can also be omitted. The final highly significant model contains the three independent variables and the interaction TENSE:VARIETY and this interaction shows the dispreference of BrE for simple past tense is significant.

In sum, Hundt and Smith's analysis has been on the right track. However, this must not distract from the fact that, especially in the case of multifactorial data sets, only a careful and statistically comprehensive analysis can ensure that all significant effects are discovered; for an example from the same journal where a more comprehensive analysis could find effects not discussed by the original authors, cf. Hommerberg and Tottie (2007) and Gries (2009b).

3.3. Further hints: useful and underutilized methods

It is impossible to provide a good overview of the many very useful aspects of the above tests and modeling methods, let alone discuss all other ways in which corpus data can be analyzed, so a few selected comments must suffice.

One important aspect has to do with non-linearity of data. Often, linear correlation measures and linear regression lines are among the first or even only statistics and graphics that are applied to data. However, a linear correlation may be significant even though the real underlying trend is not linear but curvilinear. One should therefore always plot the data – because particular relations or outliers cannot be seen from numbers alone – but also always try to summarize trends using non-parametric smoothers. These try to summarize data just like linear regression lines do, but they need not be a straight line and can therefore detect nonlinear (portions of) trends much better. One example is the dashed smoother in Figure 1.

One notion that is especially important in corpus linguistics is that of *effect sizes*. This is because the large sample sizes that many contemporary corpora provide basically guarantee that even minuscule effects will be highly significant. For example, the first analysis of Hundt and Smith's data in Section 3.2.1 returned an extremely significant result – $p<10^{-27}$ – but this is because of the sample size of nearly 160,000. The effect size, on the other hand, showed that the effect is in fact extremely small (0.03). Hence, one should always provide an effect size so that one's significant result can be better evaluated.

A second important point is concerned with Occam's razor. It is widely-known that Occam's razor dictates that non-significant predictors must be eliminated. What is less widely-known apparently is that this argument does not only apply to variables or factors, but also to levels of factors. For example, if one obtains a significant result from a table such as Table 5, then this does not mean that all four corpora differ significantly from each other. While the 'real' solution to this particular case was discussed in Section 3.2.2, such $2 \times c$ tables can also be analyzed using the so-called *Marascuilo procedure*. This procedure involves pairwise comparisons of all combinations of columns; its output can look like Table 7.

Comparisons	Differences	Critical ranges	Decision
LOB vs. FLOB	0.0013	0.0061	ns
LOB vs. Frown	0.0168	0.0058	*
LOB vs. Brown	0.0181	0.0058	*
FLOB vs. Frown	0.0155	0.0058	*
FLOB vs. Brown	0.0167	0.0058	*
Frown vs. Brown	0.0012	0.0056	ns

Table 7. Output of the Marascuilo procedure applied to Table 5

The central result is that the Marascuilo procedure shows that LOB and FLOB as well as Frown and Brown do not differ significantly from each other (and should thus be combined), but all other combinations of corpora are significantly different. This result corresponds to that of the Poisson regression from above, in which TENSE:VARIETY:TIME and TENSE:TIME were non-significant.

3.4. Exploratory methods

The final set of statistical methods to be discussed here briefly is different from most others mentioned so far. Most of the above statistics are *hypothesis-testing* in nature such that a particular *a priori* hypothesis about patterns and relations in data are tested. However, another very useful kind of methods is *exploratory* and/or *hypothesis-generating* in nature: instead of testing a hypothesis, these methods seek – in a bottom-up, data-driven fashion – patterns and relations in data sets that are too large, complex, and/or noisy to understand without such techniques. This kind of bottom-up approach is often very attractive because it relies less on any researcher's preconceptions of how the data will pattern but more on the data themselves. In fact, in corpus linguistics, where findings are often bound to be significant simply by virtue of the sample size, it is often a good idea to apply exploratory methods on top of a significance test to either test whether the variable or distinctions / variable levels tested for significance are in fact the ones most strongly supported in the data. One application of this kind was already exemplified above in Section 3.3: the Marascuilo procedure took as input four levels of the variable CORPUS and output the result that it would be better to distinguish only two kinds of corpora, which then corresponded to the variable VARIETY, whose interaction with TENSE turned out to be significant in the Poisson regression.

As usual, there is a large number of such methods: cluster analysis, principal component and factor analysis, multidimensional scaling, association rules, Bayesian classifiers, etc. The family of methods I discuss here is *hierarchical agglomerative cluster analysis*, which is one of the most versatile methods (in the sense that can handle many different kinds of input data – nominal, ordinal, interval data, raw data and distance matrices – and produces first results in the form of a tree diagram that is intuitively easy to understand. This kind of cluster analysis is typically applied to data sets consisting of n objects that are characterized by x characteristics, which may be binary, polytomous, or numeric; the main purpose is to find structure in the data set such that the n objects can be clustered into $m<n$ groups that are characterized by a large within-cluster similarity and a small between-cluster similarity.

As one example, consider Figure 3, which represents a dendrogram of nine Russian near synonyms meaning 'to try' (from Divjak and Gries 2006). The n objects clustered were the nine verbs, the x characteristics were relative fre-

quencies for 87 morphological, syntactic, and semantic characteristics of how altogether 1585 instances of these verbs were used.
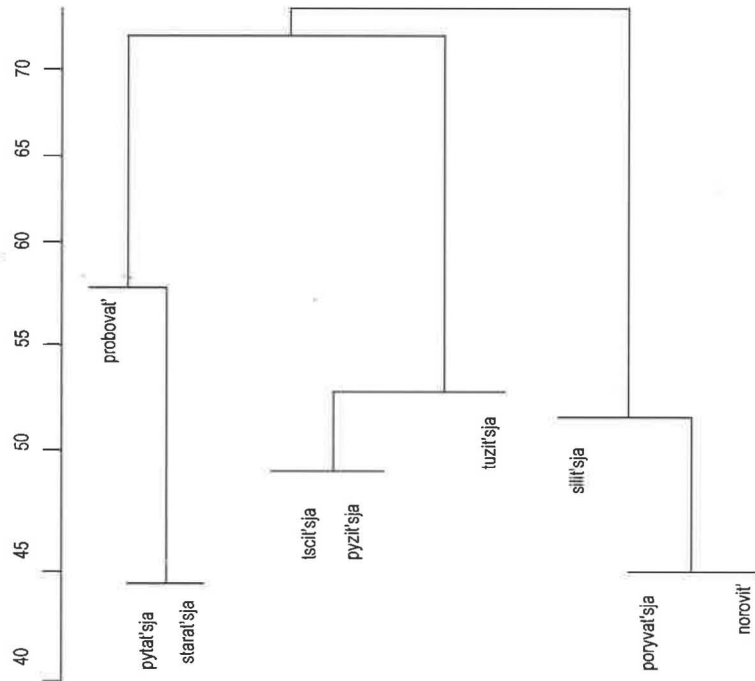

Figure 3. Clustering Russian verbs meaning 'to try'

In this case, the result is fairly clear: there are three clusters: {*silit'sja poryvat'sja norovit'*}, {*tuzit'sja tscit'sja pyzit'sja*}, and {*probovat' pytat'sja starat'sja*}, and these results can then be compared to semantic/lexicographic analyses.

While clustering methods are rather widespread in computational linguistics, their full potential has apparently not been recognized in corpus linguistics. This may be in part due to the seemingly contradictory facts that, on the one hand, cluster analyses are sometimes perceived as a mythical black box returning tree diagrams, but on the other hand, require not only statistically-informed decisions to create the tree, but also some experience and/or statistical know-how when it

comes to interpreting a dendrogram in more detail than just stating the groups or when it comes to identifying groups in more ambiguous dendrograms.

The former kind of knowledge is necessary to define (i) how the *n* objects' similarities to each other are measured and (ii) how objects that are considered very similar are amalgamated into clusters. As for (i), the analyst can choose between measures that respond to distances or curvature; as for (ii) the analyst can choose between amalgamation rules that create small even-sized or elongated clusters, but the consequences of these decisions are rather well documented in the relevant introductory literature.
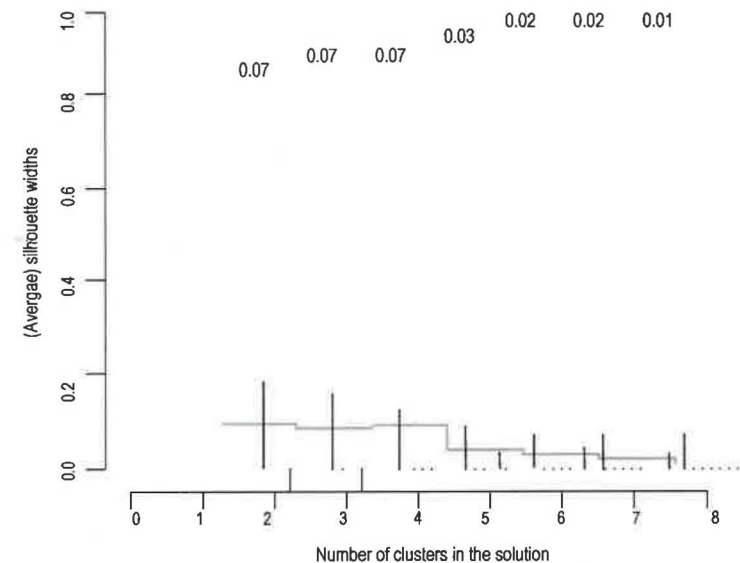

Figure 4. Average silhouette widths of different cluster solutions of Figure 3

What is much less well documented in one place is the latter kind of knowledge. One interesting way to determine the number of clusters when the dendrogram does not strongly suggest a number involves the notion of (average) silhouette widths. Simply speaking, average silhouette widths quantify the ratio between the cohesion of the elements (here, verbs) within a cluster and the cohesion of elements within a cluster to elements outside of it. The higher the average silhouette width of a cluster solution, the more discriminatory it is. This is interesting because it is then possible to compute all possible cluster solutions for a data set and determine which number of clusters results in the best discrimination. Figure 4 summarizes the results of this approach to the data of

Figure 3. The *x*-axis represents all possible and meaningful numbers of clusters for *n*=9 objects, this means testing 2, 3, ..., 7, 8 clusters), and the *y*-axis portrays the silhouette width. The black vertical lines are silhouette widths of the clusters, and the grey step function as well as the numbers at the top represent the average silhouette widths for each number of clusters. Figure 4 shows that the intuitive assessment of three clusters is confirmed.

Not only are those comprehensive tests of how many clusters there are rare to find, but often there is also no further post hoc analysis of the clustering, although these are often very interesting. For example, one can compute *F*-values for each cluster to quantify its internal homogeneity. For example, one can compute *t*-scores for each characteristic in a cluster to determine which characteristics drive the cluster solution, and there are many other approaches like these which allow for a comprehensive study of data sets. Hopefully, the many ingenuous tools statisticians have provided will find their way into suitable corpus studies.

4. Concluding remarks

Given considerations of space, the above sections could only introduce a very small number of corpus-linguistically important statistics and caveats. I hope, however, to have illustrated a few relevant methods as well as arguments why a proper use of statistical methods in indispensable, and elsewhere (e.g., Gries 2003, 2009a) I have argued and exemplified the kinds of problems that can arise from the improper statistical analysis of linguistic data. It is therefore imperative that we as corpus linguists collectively increase the level of statistical sophistication of our analyses. Next to Oakes's (1998) slightly older introduction to statistics for corpus linguistics plus some short overview articles (e.g., Biber and Jones 2009), there are now several new introductions to statistics for linguists in general available (Baayen 2008, Johnson 2008, Gries 2009b), which help familiarize the reader with incredibly powerful software.

With regard to software, it is also worth pointing out that it is no coincidence that these new introductions all use a particular piece of software for their analysis, namely the open source software R (R Development Core Team 2009). R is now the leading programming environment for statistical applications, available for different operating systems, constantly updated, and incredibly powerful in terms of range of methods, graphical exploration, sizes of data sets, etc., and can in fact also be used as full-fledged corpus-linguistic retrieval software. Given the power and the free availability of this resource, contemporary corpus linguistics should seize the opportunity and help break new ground in making our discipline empirically more precise, comprehensive, and responsible.

References

Baayen, R. H. (2008): *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*, Cambridge, Cambridge University Press.

Baroni, M. and S. Evert. (2009): "Statistical methods for corpus exploitation," in A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics: An International Handbook Vol. 2*, Berlin, New York, 777-803.

Biber, D. and J. K. Jones. (2009): "Quantitative methods in corpus linguistics," in A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics: An International Handbook Vol. 2*, Berlin / New York, 1286-1304.

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

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

Divjak, D.S. and St. Th. Gries. (2006): "Ways of trying in Russian: clustering behavioral profiles," *Corpus Linguistics and Linguistic Theory*, 2, 1, 23-60.

Evert, S. (2004): *The Statistics of Word Cooccurrences: Word Pairs and Collocations* (Unpublished doctoral dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart).

Evert, S. (2006): "How random is a corpus? The library metaphor," *Zeitschrift für Anglistik und Amerikanistik*, 54, 2, 177-190.

Evert, S. (2009a): "Rethinking corpus frequencies," paper presented at ICAME 2009.

Evert, S. (2009b): "Corpora and collocations," in A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics: An International Handbook Vol. 2*, Berlin, New York, 1212-1248.

Gale, W. and G. Sampson. (1995): "Good-Turing smoothing without tears," *Journal of Quantitative Linguistics*, 2, 217-237.

Gries, St. Th. (2003): *Multifactorial Analysis in Corpus Linguistics: The Case of Particle Placement*, London / New York, Continuum.

Gries, St. Th. (2006a): "Some proposals towards more rigorous corpus linguistics," *Zeitschrift für Anglistik und Amerikanistik*, 54, 2, 191-202.

Gries, St. Th. (2006b): "Exploring variability within and between corpora: some methodological considerations," *Corpora*, 1, 2, 109-151.

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

Gries, St. Th. (2009a): *Quantitative Corpus Linguistics with R: A Practical Introduction*, London / New York, Routledge (Taylor Francis).

Gries, St. Th. (2009b): *Statistics for Linguists with R: A Practical Introduction*. Berlin / New York, Mouton de Gruyter.

Gries, St. Th. (2009c): "Dispersions and adjusted frequencies in corpora: further explorations," in St. Th. Gries, S. Wulff, and M. Davies (eds.), *Corpus Linguistic Applications: Current Studies, New Directions*, Amsterdam, Rodopi.

Gries, St. Th. (to appear): "Methodological skills in corpus linguistics: a polemic and some pointers towards quantitative methods …"

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

Hommerberg, C. and G. Tottie. (2007): "*Try to* and *try and*? Verb complementation in British and American English," *ICAME Journal*, 31, 45-64.

Hundt, M. and N. Smith. (2009): "The present perfect in British and American English: has there been any change recently?" *ICAME Journal*, 32, 45-63.

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

Johnson, K. (2008): *Quantitative Methods in Linguistics*, Malden, MA: Blackwell.

Kita, K., K. Ogura, T. Morimoto, Y. Ueno, Y. (1993): "Automatically extracting frozen patterns from corpora using cost criteria," *Journal of Information Processing*, 34, 9, 1937-1943.

Kita, K., W. Kato, T. Omoto, and Y. Yano. (1994): "A comparative study of automatic extraction of collocations from corpora: mutual information vs. cost criteria," *Journal of Natural Language Processing*, 1, 1:21-33.

Mason, O. (1997): "The weight of words: an investigation of lexical gravity," *Proceedings of PALC'97*, 361-375.

Mason, Oliver (1999): "Parameters of collocation: the word in the centre of gravity," in J. M. Kirk (ed.), *Corpora Galore: Analyses and Techniques in Describing English*, Amsterdam, Rodopi, 267-280.

Mukherjee, J. and St. Th. Gries. (2009): "Lexical gravity across varieties of English: an ICE-based study of speech and writing in Asian Englishes," paper presented at ICAME 2009.

Nagao, M. and S. Mori. (1994): "A new method of *n*-gram statistics for large number of *n* and automatic extraction of words and phrases from large text data of Japanese," *Proceedings of the 15th Conference on Computational Linguistics*, 611-615.

Oakes, M. (1998): *Statistics for Corpus Linguistics*, Edinburgh, Edinburgh University Press.

R Development Core Team. (2009): *R: A Language and Environment for Statistical Computing*, Vienna, R Foundation for Statistical Computing, URL <http://www.R-project.org>.

Scott, M. (1997): "PC analysis of key words – and key key words," *System*, 25, 2, 233-245.

Shimohata, S., T. Sugio, and J. Nagata. (1997): "Retrieving collocations by co-occurrences and word order constraints," *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 476-481.

da Silva, J. F., G. Dias, S. Guilloré, and J. G. Pereira Lopes. (1999): "Using *LocalMaxs* Algorithm for the extraction of contiguous and non-contiguous multiword lexical units," *Proceedings of the 9th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence*, 849-849.

Wiechmann, D. (2008): "On the computation of collostruction strength: testing measures of association as expressions of lexical bias", *Corpus Linguistics and Linguistic Theory*, 4, 2, 253-290.