ZAA 54.2 (2006): 191-202 ©

STEFAN TH. GRIES

Some Proposals towards a More Rigorous Corpus Linguistics

Abstract: Over the past few decades, corpus linguistics has evolved into a fully-fledged methodological approach with an increasing number of scholars using various different methods. In this rather programmatic paper, I will argue, however, that corpus linguistics has, in some respects at least, still some way to go in terms of developing rigorous tools and methods and using them more often. More specifically, corpus linguistics – as the young discipline it still is – still has much to learn from other disciplines; a prime candidate in this respect is psycholinguistics. I will try to support this claim with arguments from several case studies.

1. Introduction

Over the past few decades, corpus linguistics has become a major methodological paradigm in applied and theoretical linguistics. Contrary to previous decades, in which acceptability judgments or grammaticality judgments made by linguists were the primary source of data, corpus-linguistic data have become more and more mainstream in general linguistics – the field of corpus linguistics is flourishing. Space limitations do not allow me to discuss in detail all the advantages that corpus linguistic methods have to offer to the linguist, but it is probably fair to say that the following is a much abbreviated and non-exhaustive list of what practitioners consider useful about corpus-linguistic work:[1]

- corpus-based quantification allows for a rather objective identification of what may be considered important and what may be considered rather marginal (of course, one may differ with respect to the frequency threshold deemed useful, but the fact that one can pinpoint a frequency threshold in terms of a number at all allows for replication etc.);
- corpus-based quantifications allows for reliable testing as well as reliability tests or comparisons of different studies more readily than studies based on subjective judgments;
- corpus-based approaches often allow for empirically more versatile studies than

[1] See McEnery and Wilson (1996, Chapter 1) and Schütze (1996) for more elaborated discussion of these and related issues.

studies based on isolated judgments;
- given the fact that corpora consist of naturally-produced speech and writing, corpus-based approaches often allow for a more valid approach than does the investigation of language produced in isolation and devoid of any context.

However, in spite of the methodological advances and the overall very promising development, many corpus-based studies exhibit a variety of what may be considered methodological shortcomings. This is particularly astonishing since, (i) presumably, for many scholars part of the reason to turn to corpus-linguistic methods may well have been a perceived dissatisfaction with methods from so-called 'armchair linguistics' (in the sense of Fillmore 1992) and (ii) there are other scientific disciplines which have already successfully coped with many of these problems. In this rather programmatic paper, I will address a few of the problems which I regard as most pressing. On a very general level, these can be divided into two groups:

- issues of how the data to be investigated are accumulated and organized;
- issue of how the data are dealt with quantitatively.

In the remainder of the paper, I will discuss problems from each group. Given space limitations, I have to restrict myself to a delineation of what I perceive as the major problematic point(s) and how I believe it/they can be overcome, but I will unfortunately not be able to exemplify my points of critique in very much detail. I will, however, repeatedly refer to studies which are directly concerned with the issue(s) at hand. Section 5 will offer a by necessity brief conclusion.

2. Where the data come from: by-subjects and by-items

The first point I wish to make is concerned with what in psycholinguistics and psychology is referred to as 'by-subjects' and 'by-items' analyses. In most experimental designs in psycholinguistics, the statistical evaluations – usually by means of analyses of variance – distinguish carefully between by-subjects statistics and by-items statistics. The former pool and average results across items and treat the experimental subjects as a random factor while the latter pool and average results across subjects and treat items as a random factor. While there is still much methodological discussion going on as to when exactly these methods are to be used in which way (cf. Clark 1973 as the most widely cited paper raising this issue), the overall objective is obvious, namely to determine to what degree the observed effects hold across subjects and items different from those actually investigated in the experiment.

Surprisingly, these methodological issues have barely found their way into corpus-linguistic studies. Many studies – and I would like to make clear from the outset that this unfortunately includes much of my own earlier work – report results for complete corpora or particular (e.g. genre-defined) parts of corpora. Such results include the frequencies of words or syntactic patterns as well as co-occurrence statistics of many different sorts, and they abound in purely descriptive work, studies from the domain of theoretical linguistics, etc. However, the point to be made below is that such a

simplification may not always be wise. The following is a case study which exemplifies this point very briefly.

The case study is concerned with the phenomenon of syntactic priming or syntactic persistence, i.e., the tendency of speakers to re-use syntactic constructions they have recently heard or produced. For example, if subjects read aloud a ditransitive sentence (the prime sentence), they are more likely to describe a picture where a person gives a book to another person with a sentence (the target) that has a ditransitive structure rather than a prepositional dative structure. There is a large body of literature on this topic from both experimental studies (cf. for instance Bock 1986 and Pickering and Branigan 1998) and corpus-based work (cf. Sankoff and Laberge 1978, Estival 1985, Szmrecsanyi 2005, Gries 2005a), and the phenomenon has been documented for native speakers of different languages (e.g., German, English, Dutch) as well as across languages (cf. Salamoura 2002, Gries and Wulff 2005). The discussion below focuses on Gries (2005a). This corpus-based study investigated syntactic persistence on the basis of two alternations, namely the dative alternation (cf. [1]) and particle placement (cf. [2]).

(1) a. *John gave his father a book.*
    b. *John gave a book to his father.*
(2) a. *John picked up the book.*
    b. *John picked the book up.*

Gries (2005a) retrieved all examples of these two construction pairs from a parsed corpus, the British component of the International Corpus of English (ICE-GB; cf. http://www.ucl.ac.uk/english-usage/ice-gb/), removed the occurrences from consideration which did not have a preceding construction (a prime) or a following construction (a target) from the pair (because then, obviously, persistence effects cannot be investigated), and coded all remaining instances (3,003 prime-target pairs instantiating the dative alternation and 1,797 prime-target pairs instantiating the particle placement alternation) for a variety of variables whose influence on persistence was investigated. The variables relevant to the present purposes include the verb lemma of the prime and the target (in [1] above both are *give*) and the file from which the constructions were taken.

The overall results, i.e. the results arrived at by simply counting and comparing all prime-target pairs across speakers and verb lemmas, showed that there are strong and highly significant persistence effects (χ^2=202.4; df=1; p<.001; Cramer's V=0.26 for the dative alternation and χ^2=183.6; df=1; p<.001; Cramer's V=0.25 for particle placement). Also, a more detailed multifactorial analysis of the effects yielded an astonishing convergence of the corpus-based results with that of earlier experimental results. However, in spite of the interesting results, what this analysis alone does not show is whether the overall results in fact mask speaker-/file-dependent results (i.e., what would correspond loosely to by-subjects statistics) and/or lemma-dependent results (i.e., what would correspond to by-items statistics), and this point of critique applies to many other corpus-linguistic studies, too.

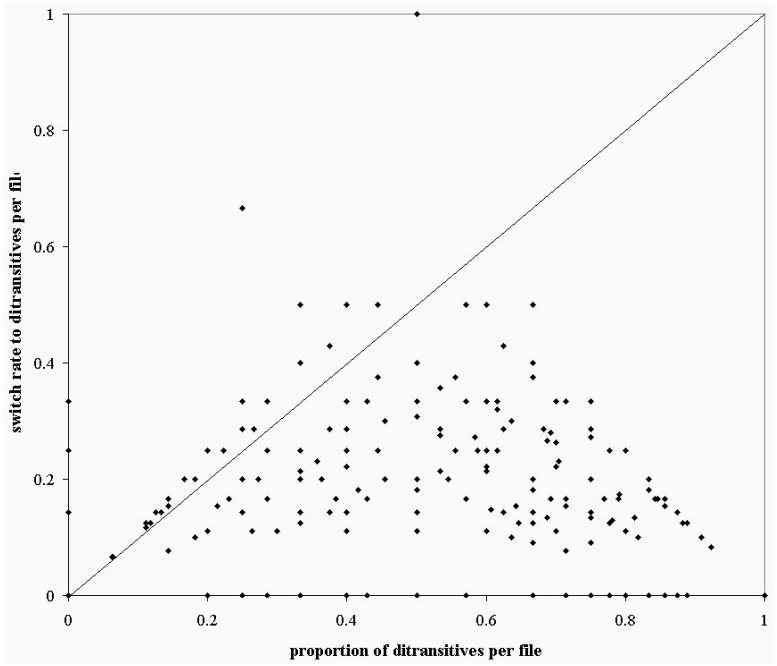

Figure 1: Switch-rate scatterplot for the ditransitive construction

One of several conceivable ways to address this issue has in fact been pointed out long ago in one of the studies on persistence cited above, namely Sankoff and Laberge (1978). They suggest using switch-rate scatterplots in which – for each speaker or file – the relative frequency of a construction on the *x*-axis is plotted against the ratio of switches to one construction on the *y*-axis. The distribution of points in the scatterplot then reveals to what degree, if any, the tendency observed in the overall analysis is also observed for the individual speakers: if the latter is true, then most of the points are below the main diagonal. Figure 1 provides the switch-rate scatterplot resulting from the analysis of the ditransitive construction in the data of Gries (2005a); the points are based on the heuristic shortcut of using files instead of speakers; the results for the other three constructions are similar in spirit (cf. Gries 2005a, Appendix A for the exact results).

For our present purposes, two aspects of these results are worth discussing. First, the switch rate observed in the individual files is much smaller than one would expect by chance: the overall persistence effect mentioned above is not called into question because it does not vary considerably or even unpredictably. However, and this is the second point, there is some variation across files – largely below the main diagonal, though – so that depending on the number and kinds of files investigated, results may in fact vary. Thus, as a first conclusion I would argue that inspecting and reporting by-subjects or by-file results should belong to the standard procedure of interpreting corpus-linguistic data.

A similar conclusion is warranted from the complementary perspective, the by-items perspective. As is obvious from the literature on syntactic persistence, there is virtually no work on the degree to which persistence may be verb-specific. While this lack of results on this issue may in part be due to the fact that most work on persistence is experimental and that it is hardly feasible to test hundreds of verbs with potentially different persistence effects, a corpus-based study allows for a much easier exploratory approach toward this issue. The question posed by the by-items statistic issue is twofold: first, one needs to find out whether individual verbs or groups of verbs are associated with particular syntactic constructions. Second, one needs to find out whether such associations influence the direction or the strength of the persistence effects.

Recent work by Gries and Stefanowitsch (2004) has been concerned with developing a method tailored to answer the first question. Their method, which is called 'distinctive collexeme analysis', is an extension of work on co-occurrence statistics of words to the co-occurrence of words with several functionally similar constructions such as the two members of the dative alternation or the two verb-particle constructions. More specifically, their method takes as input the frequencies of two functionally very similar constructions and the frequencies of words in a particular slot of these constructions (e.g., the verb slot in the ditransitive construction and the prepositional dative construction) and outputs, for each, say, verb, a statistic indicating to which of the alternating constructions this verb is attracted and from which it is repelled, and how strongly. The logic underlying the method is that the words which are most strongly associated to a particular construction are also those which indicate the semantic characteristics of a syntactic pattern or construction most strongly (cf. Gries and Stefanowitsch 2004 for motivation and exemplification).

If we now look at the data Gries and Stefanowitsch (2004) report for the ditransitive construction and the prepositional dative construction, which we are investigating presently, we find that some of the verbs used in previous experimental work differ strongly in terms of the constructions in which 'they prefer to occur':

- *show*, *offer*, and *give* are significantly attracted to the ditransitive construction;
- *sell* and *hand* are significantly attracted to the prepositional dative;
- *send* and *lend* do not exhibit a significant tendency to either construction.

In a second step, Gries (2005a) then correlated these corpus-based verb-specific preferences with the results concerning persistence and, interestingly, finds that the overall priming effect does indeed mask some strong verb-specific preferences. The verbs that are associated with one construction resist priming toward the other construction, but the verbs which do not have a significant association to either construction (as determined by the distinctive collexeme analysis) allow for priming in both directions readily.

In sum, we again find that the overall persistence effect is not called into question: speakers/writers do prefer to reuse constructions. However, the analogue to by-item sstatistics – the verb-specific investigation – revealed systematic patterns that would have gone unnoticed at the more coarse-grained level of analysis. The second

conclusion – by analogy to the first one formulated above – is that inspecting and reporting by-items results should belong to the standard procedure of interpreting corpus-linguistic data.

3. Where the data come from: the role of dispersion

A second problem I would like to mention concerns the fact that many studies investigating the use of particular words/structures have overly narrowly focused on the frequencies of these words/structures as the main diagnostic of importance. However, as I would like to point out, there may be various occasions in which it may also, or even instead, be more useful to consider the dispersion of the words/structures in the corpus under consideration.

Let me first clarify the notion of dispersion by means of an example (from Leech, Rayson and Wilson 2001): looking at word frequencies within the British National Corpus (cf. http://www.natcorp.ox.ac.uk/) they observed that the words *HIV*, *keeper*, and *lively* are all about equally frequent, namely approximately 16 occurrences per million words. While this seems to suggest that these words may be equally important in, say, a language learning context, a closer look at the results reveals that this is not the case: *HIV*, *keeper*, and *lively* occur in different numbers of corpus segments (62, 97, and 97 respectively), which indicates that – in spite of its very similar frequency of occurrence – *HIV* is a much more specialized expression than *keeper* and *lively*, an assessment that is further supported by inspecting a measure of dispersion, i.e. a measure that quantifies the degree to which a word is distributed across a corpus. Usually, high values indicate a rather even distribution while low values indicate that the occurrences of a word tend to clump together in relatively few corpus segments. For *HIV*, *keeper*, and *lively*, one such measure, Juilland's D is 0.56, 0.87, and 0.92 respectively, testifying to the assessment that the most common-or-garden word of the three is *lively*.

The above is relevant in the present context because it shows that frequency data or even, as we shall see, statistics derived from frequency data may be misleading. However, although a large variety of dispersion measures is available (the standard deviation, the variation coefficient, Juilland's D, Carroll's D_2, Rosengren's S, IDF, …; cf. for instance Oakes 1998 for an overview), most of them have not found their way into contemporary studies. In what follows, I would like to give a brief example of how looking at dispersion may not just enrich relatively simple frequency data, but also more sophisticated statistical analyses.

In Stefanowitsch and Gries (2003), the authors outline a statistical method called 'collexeme analysis' to determine which of a set of words occurring in one slot of a construction is particularly attracted to that construction.[2] Just like distinctive collexeme analysis, this method is based on an exact statistical test, the so-called

[2] Note the difference between *distinctive* collexeme analysis and (plain) collexeme analysis. The former takes words and looks at which of two constructions is preferred over the other, and how strongly – the latter takes words and identifies their attraction to / repulsion from one construction *without comparing it to another construction*.

Fisher-Yates exact test (cf. Fisher 1934, Yates 1934); and just like with distinctive collexeme analysis, the logic behind the method is, the more strongly associated a word is to a construction, the more the semantics of the word reveal about the construction. Stefanowitsch and Gries (2003) apply the method to the English imperative construction in the British component of the ICE-GB and discuss some semantic implications of the results. What is interesting in the present context is a set of verbs which all occur with a very similar frequency in the imperative in their corpus. Consider the three leftmost columns of Table 1, which provide these verbs together with their degree of attraction (positive values) or repulsion (negative values) and their frequency in the corpus.

In other words, of all the verbs that occur in the imperative between 13 and 17 times in their data, *fold* is most strongly attracted to the imperative, followed by *process*, *hang on*, and others. *Think* is most strongly repelled by the imperative, followed by *say*. Given the logic underlying the method outlined above, this implies that semantic analyses of the imperative in English should be based more on *fold* and *process* than on *hang on*, *note* or *forget*.

However, the method of collexeme analysis has one potential weakness, which is reflected in this data set. It is based on co-occurrence statistics of words and constructions alone and does *not* take into consideration how the word-construction co-occurrences are distributed within the corpus. As Stefanowitsch and Gries (2003, 237-8) briefly mention with respect to only *process* and *hang on*, this may have undesirable consequences. In fact, the slightly more exhaustive analysis here, where we include all verbs from the same frequency band, indicates that the verbs are in fact distributed across the corpus very heterogeneously. Consider the two rightmost columns of Table 1 which provide the number of files in which each verb occurs in the imperative and the resulting measure of dispersion, here Carroll's D_2.

verb	attraction/repulsion	$n_{V \text{ in imperative}}$	$n_{\text{file with V in imp}}$	Carroll's D_2
fold	21	16	1	0
process	16.7	13	1	0
hang on	16.1	17	12	0.362
note	14.5	16	10	0.342
forget	10	16	13	0.396
send	3.9	15	12	0.385
leave	2.4	17	16	0.443
write	1.7	15	10	0.354
say	-7.2	16	14	0.417
think	-7.3	15	10	0.345

Table 1: Verbs, their attraction and repulsion to the imperative, their frequency of occurrence, and their dispersion in the imperative

The result is obvious: while *fold* and *process* are more strongly attracted by the imperative than the other verbs, they occur in the imperative only in a single file whereas the other verbs, whose attraction to the imperative is slightly weaker, occur in the imperative in a much larger range of files, and these results are strongly reflected in

the measure of dispersion in the rightmost column. Obviously, relying exclusively on the frequency-based statistic of attraction/repulsion here may be misleading since this disregards the verbs' dispersion and self-evidently one would like to base one's analysis of a construction on words which are widely used in it. If one wanted to have a quantitative way of identifying which verbs are most revealing about a syntactic construction, one would need a way to simultaneously downgrade the attraction of *fold* and *process*, given their low dispersion, and upgrade the attraction/repulsion of the other verbs, given their much higher dispersion. While I am unfortunately not in a position to suggest a way how this could be done easily, I think the main message is reasonably clear: information concerning dispersion may be very useful to minimize the risk of relying too much on speakers' idiosyncrasies or (register-specific) outliers – the former establishes a connection to the topic of by-subjects statistics in Section 2 – so more care must be exercised when interpreting frequency-based data; inspecting and reporting dispersion results as a supplement to frequency data should belong to the standard procedure of interpreting corpus-linguistic data.

4. How the data are dealt with: methods and precision

Corpora can only provide frequency data, nothing else. Such data include the frequencies of morphemes, words and constructions, but of course also all kinds of co-occurrence statistics or more sophisticated statistics. While this is universally known by most corpus linguists, its implications are apparently not, namely that, if all corpus linguists have at their disposal is frequencies, the right choice of the tools with which these frequencies are evaluated becomes particularly crucial. However, even a cursory glance at recent work in corpus linguistics shows that many studies still rely on the most basic and unrefined frequency data, namely raw frequencies or percentages (cf. Atkins 1987, Berglund 1997, Boas 2003, Egan 2002, Facchinetti 2001, Hunston and Francis 2000, Johansson 2001, Kennedy 1991, Mukherjee 2003, and many more). While this is certainly not necessarily an *a priori* disadvantage, corpus linguists and computational linguists have developed a variety of statistical tools which are much more powerful and less potentially misleading than raw frequencies. Space limitations allow me only to discuss one example that is concerned with this point.

While much recent corpus-based work on the lexis-syntax interface has relied on frequency data alone, recent work by Gries, Hampe, and Schönefeld (2005, forthcoming) has provided strong empirical evidence in favour of the more sophisticated approach of collexeme analysis introduced in the previous section. The authors argue that raw frequencies or percentages have a variety of shortcomings the most crucial of which are that:

(i) raw frequencies do not allow one to identitfy the *direction* of an observed effect: is 3% of something *more* or *less* than you would expect on the basis of chance?
(ii) is the effect in whatever direction – e.g., 3% – significantly different from chance?

The authors then contrast the results of a frequency-based investigation of the *as*-predicative (cf. the pattern exemplified in [3]) with the results of a collexeme analysis

on the basis of two psycholinguistic experiments, a sentence-completion test and a self-paced reading-time study.

(3) a. $[_{VP}$ V NP_{DO} $[_{PP(?)}$ *as* $[_{XP}$]]]
b. I never saw myself as a costume designer.
c. Politicians are regarded – indeed regard themselves – as being closer to actors.

Gries, Hampe, and Schönefeld expected that speakers' preferences as to which verbs go together with an *as*-predicative are predicted better on the basis of the more refined statistical method of collexeme analysis than on the basis of the raw frequency data alone *even though the latter is still the much more widespread method*; recall the large number of references from above. This expectation was confirmed: in both experiments, collexeme analysis outperformed raw frequency data. More precisely, in the sentence-completion experiment, the effect size of the collexeme analysis was more than 60 times as high as that of the frequency analysis, and in the reading-time study, the effect size of collexeme analysis was about three times as high as that of frequency. While I cannot discuss all the results in detail here, this clearly shows that just because raw frequency data are easiest to get and process and used in most studies, there are occasions on which their results are clearly inferior to results yielded by more sophisticated methods, of which collexeme analysis is of course just one example.

While the above has shown that statistical techniques may sometimes be superior to unfiltered frequency data – which is in itself an important point – another issue remains problematic: The number of statistical tests that may be used to assess co-occurrence relations is enormous: the χ^2-test, the *t*-test, the *z*-score, Mutual Information, the binomial test, the Poisson measure, the Fisher-Yates exact test, etc. There is a large body of work by Evert and colleagues (cf. Evert and Krenn 2001) dealing with the question which of these measures are better suited to, say, collocation identification than others, but I would actually like to make the point here that this work may be problematic in a crucial respect. As a matter of fact, most of the above statistical tests are based on assumptions (of, say, normal distribution and homogeneity of variances) which natural language data usually violate. To my mind, it therefore does not make much sense to ask whether, on one particular occasion, the *t*-test yields better results than the binomial test or whether the *z*-test yields better results than the Fisher-Yates exact test. Even if the *z*-test proved superior on one occasion, the mathematical assumption underlying it would still be violated, so the question arises whether one should really use measures whose results look promising, but which are based on mathematical assumptions that are violated in one's own data. I believe the answer to this question is 'No!', which is why one should always use exact statistical tests (cf. Stefanowitsch and Gries 2003) as well as simulation and/or resampling methods (cf. Gries 2005b) wherever possible. Given the state-of-the-art in modern desktop computing, these techniques are all at our fingertips and there is no reason not to use them if we are interested in meaningful results.

A final argument in this matter is related to the previous one. It has become clear throughout the paper that I am the first to advocate rigorous statistical testing. However, one problem that often arises in corpus-linguistic studies is that, given the high frequency of some events (in some of the huge corpora currently available), many

The third line from the bottom is supposed to read
"throughout the paper that I would be the first to ..."
(Thx to Mike Scott, STG, 10/11/2007)

small findings are significant even if their practical relevance is limited (cf. Kilgarriff 2005). I would therefore like to argue in favour of doing rigorous statistical testing and providing appropriate effect sizes and considering confidence intervals as an alternative approach to significance testing proper. This way, corpus linguistics would not only avoid all the many pitfalls of null-hypothesis significance testing (cf. Loftus 1996 for insightful discussion), it would also allow the researcher to separate the wheat (perhaps barely significant but practically highly relevant results) from the chaff (highly significant but practically meaningless results).

5. Conclusions

This paper was by necessity brief and programmatic and each of the problems I mentioned would in fact deserve an article-long treatment on its own. Nevertheless, I hope to have shown that while I personally think that corpus linguistics is among the most important developments in the linguistic sciences, as corpus linguists we must also identify and be aware of the weaknesses that come with some of the methods that we apply, especially given the lively methodological discussion found in related disciplines. We need to constantly refine our methods and develop new ones with an eye to what is happening in disciplines with similarly quantitative foci: computational linguistics, psycholinguistics, psychology, etc. Only then will corpus linguistics develop into the methodological cornerstone of linguistics that many colleagues and I would like it to be.

Works Cited

Atkins, Beryl T. Sue (1987). "Semantic ID tags: corpus evidence for dictionary senses." *Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary*, 17-36.

Berglund, Ylva (1997). "Future in present-day English: corpus-based evidence on the rivalry of expressions." *ICAME Journal* 21, 7-19.

Boas, Hans Christian (2003). *A Constructional Approach to Resultatives*. Chicago, IL: The University of Chicago Press.

Bock, J. Kathryn (1986). "Syntactic Persistence in Language Production." *Cognitive Psychology* 18, 355-87.

Clark, Herbert H. (1973). "The language-as-fixed-effect fallacy: a critique of language statistics in psychological research." *Journal of Verbal Learning and Verbal Behavior* 12.4, 335-59.

Egan, Thomas. (2002). *'Let', 'allow', 'stop' and 'cease': A Corpus Study*. Paper presented at the International Conference on Construction Grammar. Helsinki, Finland, September 6 – 8, 2002.

Estival, Dominique (1985). "Syntactic priming of the passive in English." *Text* 5.1, 7-22.

Evert, Stefan and Brigitte Krenn (2001). "Methods for the qualitative evaluation of lexical association measures." *Proceedings of the 39th Annual Meeting of the ACL.* Toulouse, France, 188-95.

Facchinetti, Roberta (2001). "*Can* and *could* in contemporary British English: a study of the ICE-GB Corpus." Pam Peters, Peter Collins and Adam Smith, ed. *New*

Frontiers of Corpus Research. Proceedings from the 21st International Conference on English Language on Computerized Corpora. Amsterdam: Rodopi, 229-46.

Fillmore, Charles J. (1992). "Corpus linguistics vs. computer-aided armchair linguistics." Jan Svartvik, ed. *Directions in Corpus Linguistics*. Berlin and New York: Mouton de Gruyter, 35-60.

Fisher, Ronald A. (1934). *Statistical Methods for Research Workers*. 2nd ed. Edinburgh: Oliver and Boyd.

Gries, Stefan Th. (2003). "Testing the sub-test: a collocational-overlap analysis of English *-ic* and *-ical* adjectives." *International Journal of Corpus Linguistics* 8.1, 31-61.

Gries, Stefan Th. (2005a). "Syntactic priming: a corpus-based approach." *Journal of Psycholinguistic Research* 34.4, 365-99.

Gries, Stefan Th. (2005b). *Resampling corpora*. Paper presented at the workshop on 'Corpus statistics: Objectives, Methods, Problems'. University of Leipzig, Germany, September 29, 2005.

Gries, Stefan Th., Beate Hampe and Doris Schönefeld (2005). "Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions." *Cognitive Linguistics* 16.4, 635-76.

Gries, Stefan Th., Beate Hampe and Doris Schönefeld (forthcoming). "Converging evidence II: more on the association of verbs and constructions." John Newman and Sally Rice, eds. *Empirical and Experimental Methods in Cognitive/Functional Research*. Stanford, CA: CSLI.

Gries, Stefan Th. and Anatol Stefanowitsch (2004). "Extending collostructional analysis: a corpus-based perspective on 'alternations'." *International Journal of Corpus Linguistics* 9.1, 97-129.

Gries, Stefan Th. and Stefanie Wulff (2005). "Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora." *Annual Review of Cognitive Linguistics* 3, 182-200.

Hunston, Susan and Gill Francis (2000). *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam and Philadelphia: John Benjamins.

Johansson, Christine (2001). "Pied piping and stranding from a diachronic perspective." Pam Peters, Peter Collins, and Adam Smith, eds. *New Frontiers of Corpus Research. Proceedings from the 21st International Conference on English Language on Computerized Corpora*. Amsterdam: Rodopi, 147-62.

Kennedy, Graeme (1991). "*Between* and *through*: the company they keep and the functions they serve." Karin Aijmer and Bengt Altenberg, eds. *English Corpus Linguistics*. London: Longman, 95-110.

Kilgarriff, Adam (2005). "Language is never, ever, ever, random." *Corpus Linguistics and Linguistic Theory* 1.2, 263-76

Leech, G., P. Rayson and A. Wilson (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Longman.

Loftus, Geoffrey (1996). "Psychology will be a much better science when we change the way we analyze data." *Current Directions in Psychological Science* 5.6, 161-71.

McEnery, Tony and Andrew Wilson (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mukherjee, Joybrato (2003). "Corpus data in a usage-based cognitive grammar." Karin

Aijmer and Bengt Altenberg, eds. *The Theory and Use of Corpora*. Amsterdam: Rodopi, 85-100.
Oakes, Michael (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
Pickering, Martin J. and Holly P. Branigan (1998). "The representation of verbs: evidence from syntactic priming in language production." *Journal of Memory and Language* 39.4, 633-51.
Salamoura, Angeliki (2002). *Cross-linguistic structural priming and bilingual models of production.* Paper presented at AMLap 2002. Teneriffe, Spain, September 19–21, 2002.
Sankoff, David and Suzanne Laberge (1978). "Statistical dependence among successive occurrences of a variable in discourse." D. Sankoff, ed. *Linguistic Variation: Methods and Models*. New York: Academic Press, 119-26.
Schütze, Carson T. (1996). *The Empirical Base of Linguistics: Grammaticality Judgements and Linguistic Methodology*. Chicago, IL: The University of Chicago Press.
Stefanowitsch, Anatol and Stefan Th. Gries (2003). "Collostructions: investigating the interaction between words and constructions." *International Journal of Corpus Linguistics* 8.2, 209-43.
Szmrecsanyi, Benedikt (2005). "Language users as creatures of habit: a corpus-based analysis of persistence in spoken English." *Corpus Linguistics and Linguistic Theory* 1.1, 113-49.
Yates, Frank (1934). "Contingency tables involving small numbers and the χ^2-test." *Journal of the Royal Statistical Society, Supplement* 1, 217-35.