Reply

Helmut Küchenhoff and Hans-Jörg Schmid*

Reply to "More (old and new) misunderstandings of collostructional analysis: On Schmid & Küchenhoff" by Stefan Th. Gries

DOI 10.1515/cog-2015-0053
Received May 12, 2015; revised May 20, 2015; accepted May 22, 2015

Abstract: This is a reply to a commentary by Stefan Gries on a paper published by us in this journal in 2013. We focus on explaining the inadequacy of p-values of the Fisher Exact test as a measure of lexicogrammatical attraction and on the cognitive underpinnings of collostructional analysis. In addition, we touch more briefly on further issues, including the marginal conditioning of the Fisher Exact test and the idea of infinite strength of attraction between constructions and lexemes. We conclude with recommendations for a gold standard for applications of collostructional analysis using odds ratio, reliance and attraction scores.

Keywords: collostructional analysis, Fisher Exact test, odds ratio, attraction and reliance, entrenchment

1 Introduction

The editor-in-chief of this journal has kindly invited us to reply to Stefan Gries's commentary on our 2013 paper on collostructional analysis (CA). We gladly accept this invitation, because we think that a follow-up discussion of the issues that we have raised in our paper might be of interest to the growing community of practitioners of this method. To initiate such a discussion was our main aim in writing Schmid and Küchenhoff (2013; hence S&K 2013). While some sections of our text were quite unambiguously critical, we hoped that unbiased readers would be able to appreciate our efforts to draw attention to what lies behind CA and to balance the pros and cons of different ways of measuring lexicogrammatical associations. In

*Corresponding author: **Hans-Jörg Schmid,** Ludwig-Maximilians-Universität München, Germany, E-mail: hans-joerg.schmid@lmu.de
Helmut Küchenhoff, Ludwig-Maximilians-Universität München, Germany, E-mail: Kuechenhoff@stat.uni-muenchen.de

the "concrete recommendations" at the end of our paper, for example, and also in the intermediate summary in Section 3.6, we weighed the power and problems of different approaches, directing readers' attention not only to potential pitfalls of existing practices, but also to shortcomings of the alternatives we suggested.

Our reply falls into four main parts. Following this introduction, Helmut Küchenhoff will respond to the major statistical issues discussed in S&K (2013) and Gries's commentary in this volume. In the third section, Hans-Jörg Schmid will focus on linguistic, especially cognitive-linguistic issues. The paper ends with a short joint conclusion formulating a gold standard for future applications of CA and highlighting points of agreement between Gries and ourselves. To make it quite clear right from the start, we do not intend to reply to every single point of criticism that Gries levels, but will focus on the main issues.

2 Helmut Küchenhoff: The statistician's perspective

In my part of this joint reply to Gries's commentary I will first respond to the main statistical aspect of his criticism. In particular, I want to make clear why the use of the p-value of the Fisher Yates Exact test (FYE) as an association measure is not adequate. In my view, the odds ratio (OR) or log odds ratio (log OR) are excellent alternatives, and, if one has problems with filling the notorious cell d, so is the use of the measures of reliance and attraction. In the second part, I will address some minor points of Gries's criticism.

2.1 Measurement vs. measurement error

On the one hand, Gries repeatedly points out that there are many options for measuring the strength of collostructional associations (this issue, Section 3.1), and these have for a long time been offered as options in his R-script. On the other, however, he devotes considerable room to insisting that the p-value of FYE remains superior to other measures after all (Gries, this issue, Section 3.3.2). I cannot agree with this view. The main problem in using FYE is that it confounds two aspects of scientific measures that must not be confounded: the aspect of measuring itself, and the uncertainty about the measurement, i.e., the measurement error. This difference is of course well understood by Gries, as becomes obvious in these quotes taken from his commentary:

> ... as a significance test, [FYE] can distinguish between identical effect sizes by weighing those that are based on more data more heavily (Gries, this issue, Section 3.3.2)

> An FYE would assign more importance to a reliance value of 0.2, if it is based on 100 out of 500 rather than 1 out of 5 instances because, in the former case, the finding is likely more robust ... (Gries, this issue, Section 3.3.2).

I agree entirely with the statement that one would have a stronger trust in a result that is based on more data than in one based on less. What must definitely not be claimed, however, is that the association between lexemes and constructions is higher because it is based on more data. The only correct measure of such associations is effect size; to conflate this with the aspect of the uncertainty of the measurement in one measure is not correct. Let me illustrate this with two measurement issues in other areas.

First, suppose we want to measure the spatial distances between two pairs of equidistant points A and B, and C and D. Assume we first measure the distance between A and B with a very precise instrument which gives us a result of 1km with a possible error of +/−0.5m. One would say that the distance between A and B is 1.000km. Then we measure the distance between C and D, but this time we do not have the same instrument at our disposal, but have to make do with a much less precise measure, say a person who gives an estimate. Let us further assume that this person comes to an estimate of 1km, but due to the lacking precision of the estimate, the possible error is +/− 100m. To be sure, one would certainly "assign more importance" and would have a stronger belief in the first measurement of the distance between A and B. But would it make sense to claim that the distance between A and B is larger than or in any way different from the distance between C and D?

Second, suppose we want to evaluate the effect of two new drugs. For each drug we conduct a study comparing the drug with a placebo with regard to whether or not their application reduces the risk of contracting some health problem. In the first study, we have 100 participants, in the second study, we have 1000 participants. Assume the results were as rendered in the two parts of Table 1.

For the first drug, D1, the OR is 4, corresponding to a reduction of the risk of having health problems after therapy from 50% (25 out of 50 in the placebo control group) to 20% (10 out of 40 in the target group). The study on the second drug (D2) yields an OR of 1.5 corresponding to a risk reduction from 50% (250 out of 500) to 40% (200 out of 500). This means that the effect size in study 2 is lower than that in study 1; D2 must be credited with a less strong risk-reducing effect than D1. However, the p-values of FYE are $P_1 = 0.0031$ for the first and $P_2 = 0.0018$ for the second study. If we relied on p-values alone and thus on the conflation of effect size and sample size, as Gries recommends, then the studies would predict a stronger effect for the second drug, since P_2 is lower than P_1,

Table 1: Comparison of two fictive studies ($n_1 = 100$, $n_2 = 1000$) on the effects of two drugs using a placebo control group and a target group.

1a.

		Drug administered		Row totals
		Placebo	Drug 1	
Outcome after placebo/ drug administered	Health problem develops	25	10	35
	No health problem develops	25	40	65
	Column totals	50	50	

1b.

		Drug administered		Row totals
		Placebo	Drug 2	
Outcome after placebo/ drug administered	Health problem develops	250	200	450
	No health problem develops	250	300	550
	Column totals	500	500	

even though the actual effect size of D1 is higher than that of D2. This is clearly problematic. Transferring the two tables to the association between words and constructions would result in the conclusion that the data in Table 1(b) indicate a stronger association than those in Table 1(a). The very fact that the OR is "blind to sample size", as Gries (this issue, Section 3.3.2) correctly puts it, is in my view a necessary condition for its being a good measurement. The aspect of sample size can easily be included by giving confidence intervals for the OR. This should in fact be done, especially if the analysis is based on a small sample.

As Gries correctly states (this issue, Section 3.3.2), there is indeed a high correlation between FYE and OR. However, if that is the case then there is surely no reason not to use the OR or log OR, if it happens to be the correct measure of effect size. Note that this measurement of association has been successfully applied in thousands of medical and epidemiological studies. It has a clear interpretation, as we explained in our paper (S&K 2013: 552–555), because it measures, as its name suggests, the ratio of the odds of the co-occurrence of lexemes and constructions. This is what makes it "straightforward", as we put it, and easy to interpret. In contrast, FYE rankings are neither straightforward nor "descriptive", as Gries claims (this issue, Section 3.4.1), since they do not give effect sizes but include measure error uncertainty.

As a statistician, I would urge empirically oriented linguistic researchers to continue to apply collostructional analysis and use the R-code written and

provided by Gries, but select OR or log OR for the measurement of collostructional strength. (Just in passing: it would be wonderful if an R-package for collostructional analysis authored by Gries became freely accessible on CRAN, as is common practice in the community.) The use of p-values should be reserved for performing statistical tests when one wants to check specific hypotheses, as is of course also recommended and explained by Gries in his books.

2.2 Minor issues

2.2.1 Corrections for multiple post-hoc testing

In the context of the more or less automatic rejection of the null-hypothesis when large corpora are used, Gries discusses the value of corrections for multiple post-hoc testing. These issues are not relevant, since they pertain to the theory of testing hypotheses, while we are concerned with a measurement problem here.

2.2.2 The randomness assumption and marginal conditioning

Gries (this issue, Section 3.4.1) is right in claiming that FYE does not assume total randomness. The additional problem of FYE caused by marginal conditioning (cf. S&K 2013: Section 3.5 and Table 4), and by the fact that FYE "remains 'blind' to the internal distribution of cells nos. 2, 3 and 4" (S&K 2013: 547) are not taken up by Gries.

3 Hans-Jörg Schmid: The linguist's perspective

Like Helmut Küchenhoff, I will begin with a section on what I consider the main issue, viz. the cognitive underpinnings of collostructional analysis, and then deal with a number of minor points more briefly.

3.1 Cognitive underpinnings

Here, Gries's main concern seems to be that our discussion of the cognitive underpinnings of collostructional analysis essentially replicates ideas already

published in Stefanowitsch and Gries (2003) and especially Gries (2012), while creating the impression that it is original. This is of course a very understandable concern, which I would like to discuss from two perspectives, a temporal and a content-related one.

As far as the temporal dimension is concerned, yes, it is true that Gries (2012) was published three months before we submitted the revised version of Schmid and Küchenhoff (2013) and that we failed to take notice of it. We believe, however, that our offence is alleviated by the fact that the key elements of the discussion in our 2013 paper are based on ideas published in a paper by Schmid in 2010, which is mentioned neither in Gries (2012) nor in his commentary on our paper.

The content-related dimension is much more interesting and important, of course. Here it emerges that Gries and I have come up with highly compatible ideas on how frequency counts in corpora should be refined to elucidate better what goes on in the mind. We agree that counting and measuring the relative frequencies of co-occurring elements, say lexemes in constructions, can only give a limited amount of information on how these elements are entrenched in the associative networks of speakers. Furthermore, we agree that in addition to this type of syntagmatic relative frequency, quantitative data and statistical analyses are required on (a) the number and dispersion of lexemes occurring in the given construction, and (b) the number and dispersion of other constructions that these lexemes typically occur in, in addition to the target construction. Essentially, Gries (2012) then further argues that all these counts and calculations should be extended to data culled from several corpora, while I emphasize the importance of understanding the pragmatic aspects motivating the relative frequencies found (cf. Schmid 2014). As is typical of Gries's work, his discussion is couched in terms of numbers, scores, tables and statistical analyses and flanked by some ideas and references on implications for language learning and information processing. As is typical of my work, at least as I conceive of it, my discussion in Schmid (2010) and our account in S&K (2013) takes a complementary perspective and aims to show how entrenchment works and to what extent corpus data might be able to elucidate this process. While Gries's (2012) account mainly relies on increasingly complex exemplary contingency tables, the discussion in S&K (2013) is embedded in the wider theoretical context of the different types of associations that lie at the heart of the Entrenchment-and-Conventionalization Model (Schmid forthcoming) and builds on and extends the notions of cotext-free entrenchment, cotextual entrenchment and contextual entrenchment introduced in Schmid (2010) and further developed in S&K (2013) and Schmid (2014). All I can say about this is that I find it very encouraging to see that Gries and I have arrived at similar ideas and look forward to developing them further.

3.2 Minor issues

3.2.1 FYE (revisited) and directionality of association

As a professor of statistics, Küchenhoff naturally has to insist that measurements and measurement errors must be neither confused nor conflated. This does not mean, however, that linguists would not be in a position to accept the ranking of collexemes in terms of FYE as a useful ad-hoc measure of lexicogrammatical associations after all. If these rankings seem intuitively convincing, then why should they be rejected on purely doctrinal grounds, especially if they strongly correlate with measures of effect sizes anyway? Gries's empirical argument for the use of FYE seems to support this:

> The top collexemes of the ditransitive according to FYE are *give, tell, send, offer, show, cost, test*, ... whereas those of log odds are *give, accord, award, allocate, profit*, ... I submit that the reader, just like many audiences of talks, would find the former list more appropriate. (Gries, this issue, Section 3.4.2)

I do not think it does, however. For one thing, it should be totally irrelevant what many audiences of talks and readers of papers find more appropriate. Gries himself would be the first to emphasize that it is frequencies, scores and results of tests that count rather than intuitions. More importantly, however, he misses a point here that S&K (2013) have made very clear: it is unreasonable and futile to ponder which of the two rankings is more or less appropriate or convincing, because they give two different pieces of information. The FYE ranking is headed by verbs which are both frequent enough and sufficiently attracted by the ditransitive construction to find their way to the top of this list; in contrast, the log OR ranking reflects the verbs' propensity to occur in the construction, more or less irrespective of their absolute frequency. If a verb makes it to the top of both lists, as *give* does in Gries's example, then we can be sure that there is a very strong association.

This issue of overall and relative syntagmatic frequency is partly related to the issue of directionality, but not identical to it. Gries himself admits that "the fact that FYE is bidirectional is indeed a downside compared to other measures and I have argued for directional measures in other contexts myself" (this issue, Section 3.4.1). Yet he devotes several pages to defending FYE and insists, by means of a pointed rhetorical question, that the interpretation of FYE is "straightforward and descriptive". This claim is not only incorrect on the principled grounds detailed by Küchenhoff above, but also for the very reason that FYE represents one single dimension. If the attraction perspective from construction to lexeme and the reliance perspective from lexeme to construction are kept

apart, more information residing in the data is retained, as Gries also states (this issue, Section 3.4.1). One can assume that this insight has motivated him to include the converse one-directional measures of $\Delta P_{construction \rightarrow word}$ and $\Delta P_{word \rightarrow construction}$ as standard options in his R-scripts for the CA family. Finally, and just for the record: the obvious facts that OR shares the problem of monodimensionality with FYE and that it also requires the notorious cell no. 4 to be filled is emphasized in Schmid and Küchenhoff as many as three times (2013: 555, 557, 573–574), but Gries chooses to ignore this.

3.2.2 Empirical considerations

Gries devotes some pages to defending and extending the statistical analysis of the studies by Gries et al. (2005, 2010), which relate the results of experimental studies to corpus-based attraction measures. Admitting that the design of the studies could have been improved, he shows that for a number of very convincing reasons our re-analysis and its interpretation "can hardly be the answer" (Gries, this issue, Section 3.4.3). They were never supposed to be anyway. The main insight that emerged from our discussion was actually not related to comparing the power of different measures, which Gries aims to take some steps further in his commentary, but to show that the experimental data were not reliable enough to serve as an external benchmark in the first place (S&K 2013: 563; cf. also the note by Granvik and Taimitarha 2014: 293). In view of this, all attempts to come up with ever more sophisticated, multifactorial analyses are basically in vain, since the data on the target variable seem to be flawed. The simple observation alone that the anchor-verb *regard* "is only found at position 11 in the rank list for the experimental data" (S&K 2013: 563) is a sufficient indicator that something is fundamentally wrong with the experiment.

3.2.3 Zeros and the idea of infinitely strong attraction

Gries is of course right: the zeros in the rank lists for collostructional strength measured in terms of p-values, and the corresponding output "Inf" if the negative log is used, have nothing to do with computing power, but were due to limitations of the R-script. Apparently, the revised script published by Gries in 2014 and mentioned in his reply does not produce "0" or "Inf" any more, but is able to output precise scores for much lower, or higher if log-transformed, p-values. This script was not available when we wrote our 2013 paper. It is tempting, then, to let the matter rest and regard the new script as a happy

ending to the story of zero. However, Gries's discussion of this minor issue has started me thinking about the question what all the zeros/Infs in past applications of CA were supposed to mean.[1] Taking up Küchenhoff's clarification in Section 2 above, what they mean is that there is an infinitely small measurement error concerning the rejection of the null-hypothesis that the given lexemes and constructions are not attracted. The usual interpretation, however, is that the zeros/Infs reflect "an infinitely strong value of collostructional strength" (Hilpert 2014: 397), i.e., some sort of maximum degree of attraction. In hindsight, it turns out that it is extremely difficult to conceptualize this. To me, an infinitely strong degree of attraction could not be reached unless all occurrences of a construction are filled by all occurrences of one lexeme – an unlikely situation that would not require the application of CA anyway. In fact, some papers applying CA have even had several zeros or Infs at the top of rank lists for collostructional strength. Schönefeld (2012), for example, lists *bust*, *mad* and *wrong* with a collostructional strength of 0 for the *go*-Adj construction, Höche (2012) *get* and *happen* for the *be going to* V construction. Hilpert (2014: 396) lists no fewer than nine adjectives that are infinitely strongly attracted by the *it's* Adj *to* V-construction. How can this be conceptualized, if one keeps in mind that all these lexemes compete for co-occurrence in the same construction? The very existence of these zeros and the fact that it was not questioned what they would tell us about strengths of associations in the minds of speakers demonstrate that the implications of using FYE p-values as a measure of collostructional strength were never really considered and that CA somehow lost touch with what it was actually claiming to be a model of. Note that the OR would be "infinity" if and only if a lexeme occurs only in one construction or a construction only occurs with one lexeme.

4 Conclusion

In light of the preceding discussion, we would recommend the following gold standard for future investigations in the CA framework: if there is a theoretically sound and practically feasible way of filling cell no. 4, then the OR should be used as a unitary measure of attraction. The choice of construction counted in

1 A 15-minute search in a moderately well-stocked library unearthed more than half a dozen publications that have one or more references to zero p-values or Inf scores for collostructional strength: Backus and Mos (2012: 171, 180), Hampe (2011, 2014: 225), Hampe and Schönefeld (2012: 134), Hilpert (2014: 396), Schönefeld (2012: 28, 32) and Wulff (2006: 112).

order to fill cell no. 4 should always be justified (see Perek 2015: 123 for an example). In addition to OR, attraction and reliance scores should always be reported and interpreted as well to do justice to both directions of associations and avoid missing information available in the data (see Granvik and Taimitarha 2014 for an implementation of this standard). The identification of lexemes that are repelled by a construction (in Stefanowitsch and Gries's sense; see Gries, this issue, Section 3.4.1) is of course also possible by means of the OR, which remains below 1 in cases where the observed co-occurrence frequency is lower than the expected one. If it is impossible to fill cell no. 4, then reliance and attraction scores should be used.

Before we end, it should not go unmentioned that Gries is in agreement with a considerable number of points that S&K (2013) have addressed, among them:

- the recognition that several measures of collostructional strength can in principle be considered
- the fact that FYE values decrease automatically with growing corpus size
- the fact that the size of the score in cell no. 4 has a strong effect on the p-values calculated by the test
- the importance of making justifiable decisions about how to fill cell no. 4
- the fact that it is "definitely useful to explore [CA's] assumptions and power" (Gries, this issue, Section 4).

As pointed out in the introduction, we wrote S&K (2013) with the aim of doing precisely what Gries points out in this quote, namely to explore "theoretical premises, practical problems and cognitive underpinnings". We hope that S&K (2013), Gries's commentary and the present reply may have contributed to a better understanding of the associations between words and constructions. At the end of the day, this kind of progress is the only thing that counts in science.

References

Backus, Ad & Maria Mos. 2012. Islands of (im)productivity in corpus data and acceptability judgments: Contrasting two potentiality constructions in Dutch. In Doris Schönefeld (ed.), *Converging evidence. Methodological and theoretical issues for linguistic research*, 165–192. Amsterdam & Philadelphia: John Benjamins.

Granvik, Anton & Susanna Taimitarha. 2014. Topic-marking prepositions in Swedish: A corpus-based analysis of adpositional synonymy. *Nordic Journal of Linguistics* 37. 257–296.

Gries, Stefan Th. 2012. Frequencies, probabilities, association measures in usage-/exemplar-based linguistics: Some necessary clarifications. *Studies in Language* 36. 477–510.

Gries, Stefan Th. This volume. More (old and new) misunderstandings of collostructional analysis: On Schmid & Küchenhoff.

Gries, Stefan Th., Beate Hampe & Doris Schönefeld. 2005. Converging evidence: Bringing together experimental and corpus data on the associations of verbs and constructions. *Cognitive Linguistics* 16. 635–676.

Gries, Stefan Th., Beate Hampe & Doris Schönefeld. 2010. Converging evidence II: More on the association of verbs and constructions. In John Newman & Sally Rice (eds.), *Empirical and experimental methods in cognitive/functional research*. Stanford, CA: CSLI, 59–72.

Hampe, Beate. 2011. Discovering constructions by means of collostruction analysis: The English denominative construction. *Cognitive Linguistics* 22. 211–245.

Hampe, Beate. 2014. More on the as-predicative: Granularity issues in the description of construction networks. *Yearbook of the German Cognitive Linguistics Association* 2. 207–234.

Hampe, Beate & Doris Schönefeld. 2012. Syntactic leaps or lexical variation? More on "Creative Syntax". In Stefan Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics. Corpus-based approaches to syntax and lexis*. Berlin & New York: Mouton de Gruyter, 127–157.

Hilpert, Martin. 2014. Collostructional analysis: Measuring associations between constructions and lexical elements. In Dylan Glynn & Justyna Robinson (eds.), *Polysemy and synonymy. Corpus methods and applications in cognitive linguistics*, 391–404. Amsterdam: John Benjamins.

Höche, Silke. 2012. I am about to die vs. I am going to die: A usage-based comparison between two future-indicating constructions. In Doris Schönefeld (ed.), *Converging evidence. Methodological and theoretical issues for linguistic research*, 115–142. Amsterdam & Philadelphia: John Benjamins.

Perek, Florent. 2015. *Argument structure in usage-based construction grammar: Experimental and corpus-based perspectives*. Amsterdam & Philadelphia: John Benjamins.

Schmid, Hans-Jörg. 2010. Does frequency in text really instantiate entrenchment in the cognitive system? In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches*, 101–133. Berlin & New York: Walter de Gruyter.

Schmid, Hans-Jörg. 2014. Lexico-grammatical patterns, pragmatic associations and discourse frequency. In Thomas Herbst, Hans-Jörg Schmid & Susen Faulhaber (eds.), *Constructions – collocations – patterns*, 239–293. Berlin & Boston: Mouton de Gruyter.

Schmid, Hans-Jörg. forthcoming. A blueprint of the entrenchment-and-conventionalization model. *Yearbook of the German Cognitive Linguistic Association*.

Schmid, Hans-Jörg & Helmut Küchenhoff. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24. 531–577.

Schönefeld, Doris. 2012. *Things going unnoticed*. A usage-based analysis of *go*-constructions. In Dagmar Divjak & Stefan Th. Gries (eds.), *Frequency effects in language representation*, 11–50. Berlin & New York: De Gruyter Mouton.

Stefanowitsch, Anatol & Gries, Stefan Th. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.

Wulff, Stefanie. 2006. *Go*-V vs. *Go-and*-V in English: A case of constructional synonymy? In Stefan Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics. Corpus-based approaches to syntax and lexis*, 101–125. Berlin & New York: Mouton de Gruyter.