DE GRUYTER MOUTON
DOI 10.1515/cog-2013-0018 — Cognitive Linguistics 2013; 24(3): 531–577

Hans-Jörg Schmid and Helmut Küchenhoff

Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings

Abstract: Collostructional analysis is a corpus-based quantitative method of measuring the mutual attraction of lexemes and constructions (cf. Stefanowitsch and Gries 2003) which has gained considerable popularity among corpus linguists and especially cognitive linguists with a statistical bent. For many less statistically minded linguists, it has proven rather difficult to evaluate the theoretical background assumptions and cognitive underpinnings of collostructional analysis and to compare them to alternative ways of modelling lexicogrammatical attraction phenomena. This paper aims to spell out these premises and foundations in terms comprehensible to a wider audience. It begins with a concise survey of how collostructional analysis works and then reports on a number of practical, theoretical and statistical issues of which both practitioners of the method and those who try to appreciate results of its application should be aware. With these issues in mind we then discuss alternative ways of calculating and interpreting lexicogrammatical attraction. The advantages and disadvantages of the different methods are discussed, also against the background of the results of studies that have tried to evaluate the measures by means of external evidence from psycholinguistic experiments. Finally, cognitive underpinnings of lexicogrammatical associations and implications for the different approaches are discussed. It is argued that at present we lack adequate knowledge about the ways in which discourse frequencies affect entrenchment. We conclude that the complexities of the relation between corpus frequencies and degrees of entrenchment are still rather poorly understood, and make suggestions for future work.

Keywords: collostructional analysis, lexicogrammatical associations, quantitative linguistics, Fisher Exact test, p-value, odds ratio, attraction and reliance, entrenchment

Hans-Jörg Schmid: Ludwig Maximilians University. E-mail: hans-joerg.schmid@lmu.de
Helmut Küchenhoff: Ludwig Maximilians Universität

1 Introduction

It has been a long-standing aim of corpus linguistics to measure the degree of mutual attraction between lexical elements in text. One particularly active decade with regard to this endeavour was the 1990s, when corpora exploded in size and reliable tailor-made statistical tools were in high demand. Classic statistical procedures proposed during that period include t-score, mutual information index and log-likelihood ratio (cf. e.g. Church and Hanks 1990; Clear 1993; Stubbs 1995; Manning and Schütze 2001; see also Evert and Krenn 2001; Evert 2004). The mutual associations between lexemes and grammatical constructions (rather than other lexical elements) came into the focus of attention at the end of that period (cf. e.g. Hunston and Francis 2000; Schmid 2000), mainly because the insight was gaining ground that grammar and the lexicon are not such strictly separated modules after all. This development coincided with the first attempts within usage-based frameworks to interpret corpus-based statistical measures of associations between lexemes and patterns as reflecting degrees of cognitive associations in the minds of language users (e.g. Schmid 2000; see the survey in Glynn 2010). Descriptive measures of associations between linguistic elements thus gradually changed their theoretical status and were turned into measures of language-based associations in the minds of language users.

In 2003, Anatol Stefanowitsch and Stefan Th. Gries (henceforth S & G) introduced a set of pioneering methods subsumed under the term *collostructional analysis* (cf. S & G 2003; Gries and Stefanowitsch 2004). The major goal of these corpus-based methods is to develop improved tools for investigating interactions between lexemes and grammatical patterns. More precisely, collostructional analysis gauges the associational strength between constructions and the lexical elements filling certain slots in these constructions (S & G 2003), and unravels the semantic differences between apparently synonymous constructions ('alternations') by comparing the collostruction strength of manifestations and lexical variants in actual use as documented in corpora (Gries and Stefanowitsch 2004; see also Gries et al. 2005, 2010; Gries and Stefanowitsch 2010).

These tools have been welcomed quite enthusiastically by many members of the corpus-linguistic and cognitive-linguistic communities. Usage-based, corpus-driven, quantitative and mathematically sophisticated, collostructional analysis seems to offer a maximally objective and rigorous way of investigating not only the use of language, but also, at least if we accept the goals of usage-based approaches, degrees of entrenchment in the cognitive systems informing and guiding actual usage.

In view of the rapid spread of the collostructional methods (cf. e.g. Colleman 2009a, 2009b, 2010; Mukherjee and Gries 2009; Hilpert 2010; Hampe 2011) and the claims derived from investigations applying them, it seems important for researchers interested in the linguistic and cognitive associations of lexemes and constructions to understand the fundamental assumptions behind the method. The first aim of this paper, co-authored by a linguist and a statistician, is therefore to explain these assumptions in simple terms comprehensible to the statistical layperson. In the course of this we will point to some theoretical and practical puzzles which have so far not been brought to the attention of the wider corpus- and cognitive-linguistic community, and introduce alternative ways of measuring lexicogrammatical associations. While other researchers – e.g. Kilgarriff (2005); Divjak (2008); Ellis and Ferreira-Junior (2009); Bybee (2010); Schmid (2010); Baayen (2011); and also Gries (2005), and Gries and Stefanowitsch (2010) – have already drawn attention to some of these issues, this is the first paper to collect them in a systematic survey and to relate them to the cognitive underpinnings of measuring lexicogrammatical associations by means of collostructional analysis and other tests. The focus will be on the basic technique first introduced in S & G (2003), referred to as *collexeme analysis*, since the proposals for so-called *distinctive-collexeme analysis* made in Gries and Stefanowitsch (2004, 2010) as well as the extension proposed, for example, in Stefanowitsch and Gries (2008) largely build on the first method.

In the next section we will give a brief outline of collostructional analysis and its major premises. This will be followed by a critical appreciation of challenges faced by collostructional analysis (Section 3). In Section 4, a dataset on the N+*that*-clause construction (e.g. *the fact that* . . .) will be introduced to serve as a basis for concise accounts of alternative ways of measuring the mutual attraction of lexemes and constructions. Section 5 will review attempts to evaluate corpus-based data and different statistical tools for measuring them against evidence obtained from psycholinguistic experiments. Section 6 will provide a theoretical discussion of how corpus frequencies as such and the measures discussed relate to degrees of entrenchment of lexicogrammatical associations in the minds of speakers.

2 Collostructional analysis

As mentioned above, collostructional analysis investigates the lexicogrammatical associations between constructions and lexical elements. It is situated in the larger theoretical framework of Construction Grammar, which claims that grammatical constructions are pairings of forms and meanings, thus opposing generative models that consider grammar as a set of rules. The sizes of constructions range from individual morphemes to large-scale grammatical structures including clause-level argument-structure constructions (cf. Goldberg 1995) and grammatically exotic and lexically specific constructions such as the *let-alone* construction (Fillmore et al. 1988) or the *what's X doing Y* construction (Kay and Fillmore 1999).

The constructions investigated in collostructional analysis are typically of the syntactic type and open grammatically defined slots for lexemes to occur.

> Collostructional analysis always starts with a particular construction and investigates which lexemes are strongly attracted or repelled by a particular slot in the construction (i.e. occur more frequently or less frequently than expected). (S & G 2003: 214)

Examples studied by S & G include the ditransitive construction with a focus on the slot for the verb, the two constructions making up the so-called *dative alternation* (*she gave him the book* vs. *she gave the book to him*), and, illustrating the more specific type, the *N waiting to happen* construction with a focus on the nominal slot.

Collostructional analysis ultimately relies on frequency counts of tokens of different types of phenomena in large corpora. For successful applications of the method four different scores for frequencies of occurrence of a target lexeme (L) and a target construction (C) must be retrieved from the corpus investigated (S & G 2003: 218):

– the frequency of L in C,
– the frequency of L in all other constructions,
– the frequency of C with lexemes other than L, and
– the frequency of all other constructions with lexemes other than L.

These scores are arranged in a so-called contingency or two-by-two table familiar to many linguists from applications of the χ²-test. The setup of such tables is conventionally as rendered in Table 1:

Table 1: Contingency table cross-tabulating frequency scores of L and C

	+TARGET LEXEME	**–TARGET LEXEME**	
+TARGET CONSTRUCTION	1. frequency of L in C	3. frequency of C with lexemes other than L	row total (= frequency of C in the corpus)
–TARGET CONSTRUCTION	2. frequency of L in all other constructions in the corpus	4. frequency of all other constructions with lexemes other than L	row total
	column total (= frequency of L in the corpus)	column total	grand total

S & G (2003) illustrate this setup with data from the BNC on the noun *accident* in the construction *N waiting to happen* (cf. Table 2):

Table 2: Contingency table illustrated with data from S & G (2003: 219)

1. frequency of L in C 14	3. frequency of C with other than L 21	row total 35
2. frequency of L in all other constructions in the corpus 8,606	4. frequency of all other constructions with lexemes other than L 10,197,659	row total 10,206,265
column total 8,620	column total 10,197,680	grand total 10,206,300

More technically, and as illustrated in Table 1, the scores entered in the contingency table represent frequencies of combinations of *variables* observed in *observational units*, i.e. constructions (see Section 3.5. below for more details). The cells indicate how many observational units in the corpus exhibited the variable combinations +TARGET LEXEME +TARGET CONSTRUCTION (cell no.1), +TARGET LEXEME −TARGET CONSTRUCTION (cell no. 2), −TARGET LEXEME +TARGET CONSTRUCTION (cell no. 3) and −TARGET LEXEME −TARGET CONSTRUCTION (cell no. 4), respectively.

The contingency table serves as input to statistical tests that aim to measure the association between constructions and lexemes. Although in principle a variety of tests are available,[1] in S & G (2003) – and, as recently pointed out by Gries (2011: 240), indeed in most studies that have applied collostructional analysis – a test known as *Fisher Exact* or *Fisher-Yates* has been used.[2] The actual measure chosen to gauge the degree of attraction, referred to as *collostruction strength*, is the p-value of this test. The null hypothesis of the Fisher Exact test, as for the χ^2-test, is the independence of the occurrence of L and C. Basically, the distribution of observed frequencies is compared with expected frequencies under the null hypothesis calculated on the basis of the row and column totals, which are known as *marginals*. Given a certain distribution of observed frequencies in the corpus, the p-value indicates the probability of obtaining this distribution or a more extreme one, assuming the null hypothesis that the distribution was the result of

[1] See Wiechmann (2008) for a survey of various measures. More details on this paper will be provided in Section 5.

[2] The Fisher Exact test is part of most available statistics programmes such as R or SPSS, but it can also be found online; see Wulff (2005) for a useful survey of sites.

chance. This is interpreted by S & G as meaning that the smaller the p-value, the higher the strength of the association between lexeme and construction. More often than not, p-values are so small that their significance resides only in the number of decimal places. These scores are conventionally expressed in numbers of the type “1.12E-10” (see for example the score given for the verb *allow* in Table 3 below), which reads “1.12 times 10 to the power of minus 10”, i.e. 0.000000000112. To simplify things, a logarithmic transformation of these scores is often given,[3] which basically indicates the number of decimal places. This transformation turns the score of 1.12E-10 into “10”. Note that the larger the number of decimal places, and thus the higher the score for the logarithmic transformation, the lower the p-value, and thus the stronger the hypothetical attraction between lexeme and construction. P-values are computed individually for each of the lexemes investigated in a given construction on the basis of their observed frequencies. Once p-values have been computed for all targeted lexemes, a rank list of *collexemes* is produced, which is taken to be an indicator of the relative differences in construction strength. By way of illustration an extract provided by S & G (2003) for verbs in the ditransitive construction is rendered in Table 3:

Table 3: Rank list of collostruction strengths of top-ranking verbs in the ditransitive construction (adapted from S & G 2003: 229)

Collexeme	Raw frequency in ditransitive C	Collostruction strength
give	461	0
tell	128	1.6E-127
send	64	7.26E-68
offer	43	3.31E-49
show	49	2.23E-33
cost	20	1.12E-22
teach	15	4.32E-16
award	7	1.36E-11
allow	18	1.12E-10
lend	7	2.85E-09
deny	8	4.5E-09
owe	6	2.67E-09
promise	7	3.23E-08

[3] The negative logarithm to the base of ten of the p-value (see Gries et al. 2005: 671–672 for a discussion of advantages of this transformation).

The Fisher Exact test relates the observed ‘raw’ frequency of occurrence of a lexeme in a construction to the column and row totals. Therefore, lexemes found to occur less frequently than others in a given construction may still be found to yield a smaller p-value (and thus be more strongly attracted than more frequently found ones) if they occur less often in the corpus altogether. For example, the verb *award* ranks higher with regard to collostruction strength in the ditransitive construction than the verb *allow*, even though the latter verb occurs 18 times in the construction in the BNC, and the former no more than 7 times (S & G 2003: 229). This is because *award* is less frequent than *allow* in the whole corpus.

Table 3 also illustrates that, as is the case for the verb *give*, p-values can be so small that the computer carrying out the fairly capacity-consuming computation does not manage to give the precise score, but instead produces an output of 0. This can only be interpreted as representing a maximum degree of collostruction strength which could only be rendered more precise by using a more powerful computer.

S & G conclude their 2003 article by highlighting the major strengths of their proposal. According to them, the model increases the descriptive adequacy of grammatical description by focusing on the grammatical structures in which lexemes are embedded and by means of “the quantification of the degrees of attraction/repulsion” the method offers (2003: 236). This is seen as having positive effects on applied disciplines such as lexicography and language teaching. A second major advantage resides in the empirical support that collostructional analysis gives to construction-based syntactic theories. S & G do not fail to point out possible options for future refinements of their method, some of which have already been implemented in later papers, for example S & G (2008), Gries and Stefanowitsch (2004, 2010) as well as Gries (2006, 2011).

After this brief summary of collostructional analysis, we are now in a position to appraise the method and point to some open questions pertaining to both theoretical issues and problems arising in practical applications.

3 Critical appreciation

3.1 Null hypothesis testing and the randomness assumption

The first fundamental issue relating to the statistical side was raised by Kilgarriff (2005) in a paper emphatically entitled “Language is never, ever, ever, random”. As is indicated by the title of his paper, Kilgarriff’s main concern is the issue of randomness in linguistic data. Essentially, a random sample is a set of data in which all values are independent observations. Since there can be no doubt that all languages show distributional patterning, it is clearly problematic if we proceed from the assumption – as applications of the null hypotheses ultimately do – that the values to be observed, in our case occurrences of lexemes and constructions, are unrelated. Investigations of distributional patterns were part and parcel of behaviourist approaches in the tradition of American Structuralism (cf. e.g. Fries' [1952] method of determining word-classes), and the notions of *collocation* and *colligation* proposed by Firth, as well as Sinclair's (1991: 110) well-known *idiom principle*, precisely capture the insight that lexemes and constructions are not distributed in a random fashion (cf. Stubbs 1995: 31 *et passim*).

A second type of randomness problem, not addressed by Kilgarriff, resides in the composition of the corpora which inevitably make up the raw data of all frequency-related statistical tools. This issue can be exemplified with an analogy from the social sciences: imagine that you read the results of an opinion poll collecting 2,000 opinions on whether a given political decision was good or bad. Assessing the outcome, you would presumably be rather disappointed if you found out that the pollsters were short of informants and therefore allowed 500 persons to give four judgments each, since you would expect that each of them came to the same decision four times. Now in a way, this is almost precisely what we inevitably get in corpora: apparently, each of the language producers sampled does not contribute one datum only, for example one word, as we would expect from a proper opinion poll, but a whole stretch, or often several samples, of text. This practice, unavoidable as it clearly is, adds the problem that the observations collected in a corpus, i.e. the corpus data, are not randomly sampled. Speakers and writers have their favourite ways of putting things, habitually resort to the same fixed phrase and collocations and frequently reproduce identical chunks of text very much in the fashion of ready-made building-blocks (cf. again Sinclair's *idiom principle*, 1991: 110, as well as Szmrecsanyi 2005). As a result, the phenomena collected in a corpus can never be 'independent observations'. To be fair, it must be emphasized that this problem is by no means specific to collostructional analysis, but affects corpus-linguistic practice and theory per se. It is aggravated, however, if statistical tests are used which start out from a null hypothesis and are therefore based on the assumption of independent observations.[4]

[4] In principle, statistical methods for getting away from the assumption of independence could be used, for example, mixed models including random effects, e.g. for speakers and sources, but it is not clear how these could be applied in order to improve measures of lexicogrammatical associations.

3.2 P-value of a significance test as a measure of collostruction strength and resulting problems with the interpretability of scores

The output of collostructional analysis consists of lexemes ranked according to p-values, which are interpreted as indicating different degrees of collostruction strength. As explained in Section 2, the rationale behind this is essentially that the Fisher Exact test measures, in the form of the p-value, the probability that the distribution actually observed, or a more extreme one, occurs if there is no attraction between the lexeme and the construction. It is important to understand what this means. The p-value is a measure of the evidence of a set of data with regard to a certain hypothesis. The lower the p-value, the stronger the evidence against the null hypothesis. What the p-value does not do, however, is measure the strength of a relation, be it lexicogrammatical or other. As Baayen (2011: 16) observes with reference to collostructional analysis: "From a statistical perspective, it is somewhat odd to derive a measure from a p-value".[5] S & G are of course also aware of the difference between p-values and effect sizes and explicitly note that "ranking the lexemes [. . .] would normally have to be done using effect sizes" (S & G 2003: 239). Justifying their choice of p-values to measure collostruction strength, they add that

> the advantage of the Fisher exact p-value is that in addition to incorporating the size of the effect observed in any particular cross-tabulation (as, e.g., Φ, MI or the odd's ratio would also do), it also weighs the effect on the basis of the observed frequencies such that a particular attraction (or repulsion, for that matter) is considered more noteworthy if it is observed for a greater number of occurrences of the lexeme in the N slot. (S & G 2003: 239)

While it is not quite clear in which way the Fisher Exact p-value indeed, as S & G put it, "incorporat[es] the size of the effect", in a later publication the authors point out that "alternative measures such as effect sizes [. . .] could also be used" (S & G 2009: 943). What this shows, and what should be kept in mind in interpreting rank lists of lexemes ordered according to p-values, is that p-values do not, strictly speaking, measure the strength of the association between lexemes and constructions, but rather the likelihood with which the assumption that there is

[5] In the statistical literature it is widely accepted that p-values must not be seen as an effect measure, see e.g. Goodman (2002: 593): "Because the p value is calculated only with respect to one hypothesis, and has no information, by itself, of the magnitude of the observed effect (or equivalently of power), it implicitly excludes the magnitude of effect from the definition of 'evidence'."

no attraction, i.e. the null hypothesis, can be rejected. As a consequence of this way of operationalizing the measure of collostruction strength, the interpretability – in a technical, statistical sense – of rank orders is more restrained and less transparent than that of actual effect sizes.

3.3 Sample size

The next issue is the dependence of the p-value of the Fisher Exact test on the sample size, which is also mentioned in the quotation from S & G (2003) rendered in the preceding section. Theoretically, of course, it is clearly reasonable to assume that observed frequencies in very large corpora are treated as being more informative than data collected from smaller corpora. An observed frequency of, say, 10 records of a phenomenon in a corpus of 1 million words is clearly a less reliable datum than an observed frequency of 1,000 in 100 million words, even though the relative frequency is 10 per million words in each case. However, simply due to their sheer size, the large corpora available today have an in-built potential to reject the null hypotheses more or less automatically. Kilgarriff gives an interesting quote from a statistics textbook from the 1970s:

> None of the null hypotheses we have considered with respect to goodness of fit can be *exactly* true, so if we increase the sample size (and hence the value of χ^2) we would ultimately reach the point when all null hypotheses would be rejected. All that the χ^2 test can tell us, then, is that the sample size is too small to reject the null hypotheses! (Owen and Jones 1977: 359, quoted from Kilgarriff 2005: 266)[6]

If the sample size increases, then the p-value decreases, even if the internal structure of the dataset remains unchanged. For example, when the numbers in the two-by-two contingency table are all doubled, then the p-value decreases, even though one would assume that the attraction strength remains constant, as the proportions between the numbers remain constant, too. Well aware of this, Gries (2005) emphasizes that comparisons of p-values must always be based on identical corpus sizes. One way to react to the sample size problem, which is also pointed out by S & G (2009: 943), would be to replace the Fisher Exact test by a different distributional statistic that is not affected by sample sizes, e.g. the Odds Ratio measure, which will be explained in section 4.3 below.

[6] This danger has been acknowledged in a reply to Kilgarriff's paper by Gries (2005), who adds a substantial list of further warnings related to the frequency effects of null hypothesis testing.

3.4 The challenge of filling cell no. 4 (as well as the other three cells)

As explained in Section 2, like most statistical computations, collostructional analysis compares observed frequencies of occurrence in a corpus with frequency distributions that would be expected by chance, i.e. when the null hypothesis is correct. Expected frequencies are calculated on the basis of the row and column totals as illustrated in Table 1. These totals, as well as the grand total, can of course only be calculated if all four cells of the contingency table are filled. However, retrieving the scores needed to fill all four cells is by no means a trivial task. The challenges facing researchers here are definitorial ones. From a statistical point of view, they concern the definitions of the *observational unit* under examination and of the *variables* to be investigated as well as their values. As mentioned in Section 2, in collostructional analysis, the variables are represented by the target lexemes and the target construction, both of which are binary variables and thus have two values (+TARGET LEXEME VS. –TARGET LEXEME and +TARGET CONSTRUCTION VS. –TARGET CONSTRUCTION). The observational unit under examination is commonly formulated by selecting a more schematic construction. All of these definitions and choices deserve closer examination.

Firstly, a clear definition of the given target lexeme, with regard to both its form(s) and its meaning(s), is a prerequisite for collecting the scores for the cells numbered 1 and 2 in Table 1. This definition is the basis for counting the frequency of occurrence of the target lexeme in the target construction (cell no. 1) and the frequency of the same lexeme in all other constructions (cell no. 2), or, put more technically, the number of constructions (*qua* observational units) that contain the target lexeme and contain or represent the target construction (cell no. 1) and the number of constructions that contain the lexeme but do not contain or represent the target construction (cell no. 2). Decisions that have to be made in the course of this firstly relate to the question as to whether all morphological variants of a lexeme are included in the count or whether they are counted separately (cf. Gries 2011). Furthermore, one should be aware that, strictly speaking, it is not forms that enter into lexicogrammatical associations, but *lexemes* (*qua* abstract bundles of meanings), or even more precisely *lexical units* (*qua* associations of forms and senses; cf. Lipka 2002: 150 for these terms). It is important to emphasize this, since what all corpus linguists – not only those applying collostructional analysis – usually do when filling cell no. 2 is count forms. The reason for this lies in the amount of material to be processed and the effort required in order to handle it properly. While it is time-consuming enough to check individual tokens manually when it comes to filling cell no. 1, a semantically informed way of filling cell no. 2 would go way beyond that in terms of time and effort, as it would entail a manual inspection of **all tokens of all** collexemes of a construction. Most linguists will certainly agree that this does not seem to be feasible in most studies for practical reasons (cf. Stefanowitsch and Gries 2008: 149). Nevertheless, it would not be unproblematic from a semantic (and also cognitive) point of view if, for example, light-verb uses of *give* (e.g. *give sb responsibility, give sb a smile, give sb confidence,* etc.) were included in the count of all uses of *give* in the corpus when investigating the ditransitive construction, or if uses of *I see* in discourse marker function ('I understand') were included in the count of all uses of *see* in a corpus when investigating the *as*-predicative construction (e.g. *regard as, view as,* etc., Gries et al. 2005; see Section 5 for more details).

Secondly, the score to be inserted in cell no. 1 of course also depends on the definition of the second variable, the construction under examination. This definition is also necessary for cell no. 3, which represents the number of constructions containing or representing the target construction but not the target lexeme (+TARGET CONSTRUCTION, –TARGET LEXEME). As in the case of the target lexemes, an exact definition of the target construction with regard to its formal and semantic properties is required, which is often even more difficult to work out since constructions are even messier and more flexible than individual lexemes. The formal description includes a precise account of the forms and functions of fixed lexical and grammatical elements that define the construction. The semantic description must detail the meaning of the target construction. Once a definition is in place, its application to the corpus data will usually have to be carried out manually or semi-manually, depending on the amount of annotation added to the corpus. Again, this is a problem that all attempts to measure lexicogrammatical attraction phenomena have to grapple with. The corpus analysis can turn out to be particularly difficult for potentially polysemous constructions such as the *as*-predicative construction investigated by Gries et al. (2005, 2010; cf. Section 5). For example, while *regard as, see as* and *view as* seem to be representatives of the core meaning of this construction, *use as* and *offer as* are both grammatically and semantically fairly distinct: the verbs *use* and *offer* can be used perfectly well without an object complement and do not seem to have the epistemic meaning associated with the core sense of the construction. In order to work out how to fill cell no. 3, a decision has to be made as to whether the latter group of verbs in fact instantiate the construction and are thus to be included in the count, or are to be treated as a separate sense of the target construction or even a semantically distinct homonymous construction, which may only be related to the target construction on a much higher level of schematicity.

Thirdly, the cell which is the most difficult one to fill is the one numbered 4 in Table 1 above. Defined rather loosely as rendering "the frequency of all other constructions with lexemes other than L", i.e. the target lexeme (S & G 2003: 218), cell no. 4 actually brings to the fore the knotty question of how to define the statistical or observational unit under examination. S & G explicitly state how they deal with this question in connection with the construction *N waiting to happen* (2003: 218):

> the total number of constructions was arrived at by counting the total number of verb tags in the BNC, as we are dealing with a clause-level construction centering around the verb *wait*.

As is implied in the second part of this quotation,[7] the score in cell no. 4 should meet two criteria: it must render the number of constructions in the corpus which feature the value intersection -TARGET LEXEME and -TARGET CONSTRUCTION, and it should also be derived from the total number of constructions which are defined in such a way that they are **somehow comparable** to the target construction. In technical parlance, comparability is among the key *inclusion criteria* in the definition of the observational unit, which also determine the size of the total population studied. More specifically, if the scores in the two-by-two contingency table and the application of the Fisher Exact – or any other statistic based on the table – are to make sense both mathematically and linguistically, the target construction investigated, i.e. *N waiting to happen*, should be a subset of the set of constructions defined as observational units under examination, i.e. all verbs, or, more precisely, all verbal constructions. This implies that the two should be paradigmatically related in terms of their forms, functions and meanings (see Section 6 for more details). S & G's choice of "verb tags" to select the paradigmatic competitors of the *N waiting to happen* construction seems plausible enough, but it is of course not uncontroversial (cf. Bybee 2010: 98). One could argue, for example, that the construction is lexically so specific, including as it does the verbs *wait* and *happen*, that it seems unfair, so to speak, to relate it paradigmatically to all main-verb constructions. In addition, or alternatively, one could point to the fact that the verb *wait* invariably occurs in the progressive form and conclude that only progressive verb phrases should be taken into consideration (cf. Gries 2011). A similar argument could be constructed for the infinitival form *to happen*, which would lead to an even smaller score for cell no. 4.

Note that the decisions to take when filling cell no. 4 are not just practical ones, but rather pertain to very fundamental questions of construction grammar.

[7] Cf. also Gries et al. (2005: 645): "Fourth, one estimates the number of constructions in the corpus [. . .]. For the analysis of argument structure constructions we have adopted the strategy advocated in the first works on collostructional analysis, namely to approximate this frequency by using the token frequency of all verbs."

Firstly, the decision concerns the definition of the nature and size of the construction serving as observational unit. Since the extent of constructions can range from simple morphemes to complex argument-structure constructions, the ways in which target constructions contain, instantiate or represent the type of construction defined as observational unit must be clarified. This, secondly, involves the allocation of a place for both the target construction and the construction defined as observational unit in the network of the myriads of constructions conventionalized in a given language. Thirdly, and more specifically, the decision how to fill cell no. 4 relates to the nature of the construction's links to other constructions, mainly in terms of their schematicity (cf. Zeschel 2009; Gries 2011). Applied to the *N waiting to happen* construction, we would presumably have to select the immediately 'superordinate' construction from a range of more or less specific or schematic candidates: is the construction a more specific variant of the V-*to*-V-construction, or of the V_{progr}-to-V-construction, or, as S & G suggest, simply of the most schematic main-verb-construction?

The score entered in cell no. 4, like the ones filling cells nos. 2 and 3, thus depends on subjective decisions made on the basis of linguistic theorizing. These decisions have far-reaching consequences for the outcome of Fisher Exact because the size of the score in cell no. 4 has a strong effect on the p-values calculated by the test. The larger the score entered in cell no. 4, the larger the row and column totals, which are part of the formula for calculating the Fisher Exact p-values, and thus the smaller the p-values, as long as the proportions in the contingency table remain constant (cf. Section 3.3).[8]

The cell no. 4 problem forces a choice upon researchers, not only upon those applying collostructional analysis, but indeed upon all researchers who work with contingency tables derived from corpus frequencies. All inevitably have to weigh the difficulty of coming up with linguistically sound and mathematically feasible ways of filling this cell against the need to take the number of other observations in the corpus into account in order to render statistical measures valid and reliable. As we will see in Section 4, only the first alternative approach suggested in this paper, the Attraction and Reliance approach proposed in section 4.1, does not have to grapple with the cell no. 4 problem, while the second and the third ones, Delta P (Section 4.2) and Odds Ratio (Section 4.3), are indeed confronted with it, too, and thus do not provide a solution.

[8] Bybee (2010: 97) claims that "high overall token frequency of a lexeme detracts from its Collostructional Strength". This is not confirmed by our calculations (see also Section 3.5). The difference in results may have to do with the fact that Bybee does not seem to be concerned with the relation of the "overall token frequency of a lexeme" to the other marginals of the two-by-two contingency table, which determine the size of the p-value of Fisher Exact.

3.5 Directionality of association, marginal conditioning of Fisher Exact and effects of high frequencies of lexemes outside target constructions

Collostruction strength, as operationalized by p-values of the Fisher Exact test, was originally introduced to measure “the interaction of lexemes and the grammatical constructions associated with them” or, more precisely, to show how lexemes “can [. . .] be ranked according to their strength of association (the Fisher exact p-values, that is) with the construction” (S & G 2003: 219). This seems to imply that the two-way association between constructions and lexemes, on the one hand, and lexemes and constructions, on the other, can be captured by one single measure. However, Ellis and Ferreira-Junior (2009) have made it clear that such “associations are not necessarily reciprocal in strength” and that “[t]hese directional relations therefore need to be separately assessed” (2009: 198). From one perspective, we treat the construction as given and examine which lexemes are attracted by it; the reciprocal perspective proceeds from a given lexeme and looks at the constructions in which it is found. In two-by-two contingency tables, the construction-based perspective corresponds to the horizontal examination of the first row, and the lexeme-based perspective to the vertical examination of the first column (see Section 4.2 for more details). Since the output of collostructional analysis is a ranking in terms of only one score, the p-value of Fisher Exact, the measure is unable to tease apart these two perspectives. Fundamental differences that emerge only when both directions of associations are examined separately can be levelled out by p-values as a consequence. Consider the two fictive datasets presented in Table 4. On the left-hand side, in Table 4a, the two-by-two table of a very rare lexeme is depicted which yields no more than 100 tokens in a 10-million-word corpus, of which as many as 40% are found in the construction investigated. On the right-hand side, in Table 4b, we see fictive data for a much more frequent lexeme which shows a yield of 119 tokens in the target construction, as compared to 4,881 occurrences in other constructions. The scores are manipulated in such a way that the row totals and the grand total are the same in both cases, which demonstrates that both “belong” to the same construction in the same corpus. From the construction-based perspective, it can be observed that the lexeme in 4b co-occurs more frequently with the construction than the lexeme in 4a (119 vs. 40). From the lexeme-related perspective, however, the data tell us that the lexeme in 4a co-occurs with the construction relatively much more frequently than the lexeme in 4b (40% vs. 2%). Metaphorically speaking, the lexeme in 4a depends on the construction to a much greater extent than the one

Table 4: Juxtaposition of two fictive frequency distributions and their p-values

4a)

1. frequency of L in C 40	3. requency of C with other than L 21,960	row total 22,000
2. frequency of L in all other constructions in the corpus 60	4. frequency of all other constructions with lexemes other than L 9,977,940	row total 9,978,000
column total 100	column total 9,999,900	grand total 10,000,000

Fisher Exact, p-value = 6.64 E-79

4b)

1. frequency of L in C 119	3. frequency of C with other than L 21,881	row total 22,000
2. frequency of L in all other constructions in the corpus 4,881	4. frequency of all other constructions with lexemes other than L 9,973,119	row total 9,978,000
column total 5,000	column total 9,995,000	grand total 10,000,000

Fisher Exact, p-value = 5.81 E-79

in 4b does, even though the lexeme in 4b is attracted more frequently by the construction.

The point of these two scenarios, which can be gleaned from the line at the bottom of the table, is that despite the very disparate sets of scores, the p-values are essentially the same, indicating that the collostruction strengths of the two lexemes in the construction have to be considered to be practically identical. The reason for this lies in the fact that the Fisher Exact test relies only on cell no. 1 – rendering the occurrences of a lexeme in the construction – and the marginals. It does not take into account the relations between cell no. 1 and cell no. 2, on the one hand, and cell no. 1 and cell no. 3 on the other. Since Fisher Exact remains “blind” to the internal distribution of cells nos. 2, 3 and 4, p-values run the risk of conflating different types of associations between lexemes and constructions which should presumably be kept apart. As the fictive scenarios in 4a and 4b show, the effect of the marginal conditioning of the test is particularly strong when the score in cell no. 2, rendering occurrences of the given lexeme in other constructions, is very high, because this creates a high marginal in the first column.

3.6 Summary

To provide an intermediate summary, we can conclude that the following issues have to be kept in mind in applications of collostructional analysis and interpretations of the results that they produce:

1. Null-hypothesis testing is based on the assumption that the data are distributed randomly, which is presumably not the case for corpus data.
2. The observations in a corpus do not meet the requirements of random data sampling.
3. The use of a significance measure (the p-value of the Fisher Exact test) as a measure of collostruction strength causes problems for interpretability.
4. Larger samples reduce p-values as compared to smaller samples with the same internal distribution.
5. A range of theoretical and practical issues have to be taken into consideration when determining the scores to be entered in contingency tables, especially the score filling cell no. 4.
6. The directionality – from construction to lexeme vs. lexeme to construction – should be kept in mind.
7. The Fisher-Exact test is conditioned on the marginal distributions in the contingency table.
8. High absolute frequencies of lexemes outside the target construction affect p-values.

It should be emphasized that the first and the second points are by no means specific to collostructional analysis, but rather are shared by a wide range of well-established corpus-linguistic statistics. S & G themselves have stressed that the use of the Fisher Exact test and its p-values is not the only option, though it has clearly been used most frequently in existing applications of collostructional analysis. One further point of criticism, levelled by Bybee (2010: 98), is that "[p]roponents of Collostructional Analysis hope to arrive at a semantic analysis, but do not include any semantic factors in their method". Although we are not sure whether Bybee's additional claim that "[s]ince no semantic considerations go into the analysis, it seems plausible that no semantic analysis can emerge from it" is actually correct, we will come back to this issue in Section 6.

In the next section, we will discuss three alternative approaches to measuring the mutual attraction of constructions and lexemes and assess them with regard to their potential to tackle the challenges faced by collostructional analysis.

4 Alternative approaches

To be able to illustrate the following discussion with examples, we introduce a dataset on a nominal construction, the N+*that*-clause construction (e.g. *the fact that . . .* , *the news that . . .* ; cf. Schmid 2000). The data are extracted from the International Corpus of English – Great Britain (ICE-GB), which contains only one million words and thus avoids the problem of infinite, i.e. zero, scores for collostruction strength encountered by Schmid (2010). A second advantage of ICE-GB is that it is fully parsed and therefore allows data retrieval with a good ratio of recall and precision. Since the data is mainly used for illustrative purposes, only the 40 nouns occurring most frequently in the construction will be listed (cf. Table 5). These 40 types account for approximately 83% of all valid tokens of the construction. The query <N(com,sing)> that+<CONJUNC> yielded 1,149 hits; manual post-processing identified 601 unwanted hits, leaving us with 548 valid hits. Table 5 lists the scores related to this dataset for p-values of the Fisher Exact test as well as the three measures discussed in this section – Attraction and Reliance (Section 4.1), Delta P (ΔP; Section 4.2) and Odds Ratio (Section 4.3). For those tests that require cell no. 4 in the contingency table to be filled, it should be known that the set of singular nouns was chosen as the observational unit under examination, amounting to 111,387 tokens in the BNC.

4.1 Attraction and Reliance

The measures of *Attraction* and *Reliance* were proposed prior to the advent of collostructional analysis by Schmid (2000: 54–55). The idea of the two measures is to do justice to the directionality issue (see Section 3.5) by separating the proportion with which a grammatical construction is filled by a given lexeme from the proportion with which a lexeme occurs in a given construction. The former proportion is interpreted metaphorically as reflecting the *Attraction* exerted by the construction on the lexeme, the latter as reflecting the *Reliance* of the lexeme on the construction. *Attraction* is calculated by dividing the frequency of occurrence of a noun in a construction by the frequency of the construction in the corpus; *Reliance* is calculated by dividing the frequency of occurrence of a noun in a construction by its frequency of occurrence in the whole corpus.[9] To be able to render the scores as percentages, the dividend is multiplied by 100 in both divisions. The calculation is shown in Table 6 and illustrated using the scores for the nouns *suspicion* and *sign* given in Table 5:

The scores indicate that the noun *suspicion* accounted for 0.73% of the uses of the N-*that* construction in the ICE-GB, and that 30.77% of the uses of the same noun were found in the N-*that* construction. The noun was thus attracted with a

[9] The measure of Attraction thus corresponds to relative frequency in the construction. What is captured by the measure of Reliance has been referred to as *Faith* in later publications by Gries and others (see Section 5)

Table 5: Nouns attracted by the *N-that* construction with data from ICE-GB (ranked according to Reliance)

noun	freq. in pattern	freq. in corpus	p-value Fisher Exact	Attrac-tion	Reliance	ΔP Reliance	ΔP Attrac-tion	Odds Ratio
fact	161	251	6.21E-314	29.38%	64.14%	0.6379	0.2930	511.93
assurance	6	13	2.30E-11	1.09%	46.15%	0.4567	0.0109	175.27
assumption	11	24	8.71E-20	2.01%	45.83%	0.4535	0.0200	174.63
realis/zation	4	9	7.16E-08	0.73%	44.44%	0.4396	0.0073	162.99
suggestion	11	33	6.49E-18	2.01%	33.33%	0.3285	0.0199	103.18
suspicion	4	13	4.00E-07	0.73%	30.77%	0.3028	0.0072	90.55
probability	6	23	1.30E-09	1.09%	26.09%	0.2560	0.0108	72.17
impression	13	52	4.58E-19	2.37%	25.00%	0.2452	0.0234	69.03
proposition	3	13	3.27E-05	0.55%	23.08%	0.2259	0.0054	61.01
doubt	25	113	5.96E-34	4.56%	22.12%	0.2166	0.0448	60.16
belief	7	32	2.03E-10	1.28%	21.88%	0.2139	0.0125	57.35
expectation	5	23	8.85E-08	0.91%	21.74%	0.2125	0.0090	56.69
proof	5	23	8.85E-08	0.91%	21.74%	0.2125	0.0090	56.69
indication	6	31	9.15E-09	1.09%	19.35%	0.1887	0.0107	49.07
notion	6	31	9.15E-09	1.09%	19.35%	0.1887	0.0107	49.07
confirmation	3	16	6.32E-05	0.55%	18.75%	0.1826	0.0054	46.93
hope	10	55	1.84E-13	1.82%	18.18%	0.1770	0.0178	45.76
feeling	15	83	1.77E-19	2.74%	18.07%	0.1760	0.0268	45.84
speculation	3	18	9.15E-05	0.55%	16.67%	0.1618	0.0053	40.67
conclusion	5	34	7.00E-07	0.91%	14.71%	0.1422	0.0089	35.18
sign	6	60	5.51E-07	1.09%	10.00%	0.0952	0.0105	22.71
recognition	3	32	0.0005282	0.55%	9.38%	0.0889	0.0052	21.03
possibility	8	88	1.48E-08	1.46%	9.09%	0.0861	0.0139	20.51
evidence	26	287	1.02E-24	4.74%	9.06%	0.0859	0.0451	21.1
view	24	273	1.32E-22	4.38%	8.79%	0.0832	0.0415	20.34
idea	24	297	9.72E-22	4.38%	8.08%	0.0761	0.0413	18.55
fear	5	64	1.70E-05	0.91%	7.81%	0.0733	0.0086	17.29
understanding	4	52	0.0001301	0.73%	7.69%	0.0721	0.0069	16.97
claim	6	80	3.05E-06	1.09%	7.50%	0.0702	0.0103	16.57
danger	4	59	0.0002128	0.73%	6.78%	0.0629	0.0068	14.81
thought	4	61	0.0002422	0.73%	6.56%	0.0607	0.0068	14.29
principle	4	65	0.0003093	0.73%	6.15%	0.0567	0.0067	13.35
knowledge	7	114	1.75E-06	1.28%	6.14%	0.0566	0.0118	13.39
concern	3	58	0.0029893	0.55%	5.17%	0.0469	0.0050	11.09
risk	3	76	0.006381	0.55%	3.95%	0.0346	0.0048	8.35
statement	4	121	0.0031272	0.73%	3.31%	0.0282	0.0062	6.96
news	6	201	0.0005222	1.09%	2.99%	0.0250	0.0092	6.28
theory	4	135	0.0046166	0.73%	2.96%	0.0248	0.0061	6.21
issue	3	156	0.0423285	0.55%	1.92%	0.0144	0.0041	3.98
problem	3	372	0.4364515	0.55%	0.81%	0.0032	0.0021	1.65

Table 6: Calculating Attraction and Reliance scores

$\text{Attraction} = \frac{\text{cell } 1 \times 100}{\text{cell } 1 + \text{cell } 3}$	$\text{Attraction}_{\textit{suspicion that}} = \frac{4 \times 100}{548} = 0.73\%$
	$\text{Attraction}_{\textit{sign that}} = \frac{6 \times 100}{548} = 1.09\%$
$\text{Reliance} = \frac{\text{cell } 1 \times 100}{\text{cell } 1 + \text{cell } 2}$	$\text{Reliance}_{\textit{suspicion that}} = \frac{4 \times 100}{13} = 30.77\%$
	$\text{Reliance}_{\textit{sign that}} = \frac{6 \times 100}{60} = 10.00\%$

proportion of 0.73% by the construction and relied on the construction with a proportion of 30.77%. In contrast, the noun *sign* is a slightly more important slot filler for the construction (Attraction score 1.09%), but relies on the construction to a considerably lesser extent (Reliance score 10%). In spite of these differences, the two nouns have practically identical p-values of 4.00E-07 and 5.51E-07 respectively, which again illustrates the problem discussed in Section 3.5. In terms of two-by-two contingency tables, Attraction is calculated by dividing cell no. 1 (occurrences of L in C) by the row total (cell 1 + cell 3; all occurrences of C), while Reliance is the division of cell no. 1 and the column total (cell 1 + cell 2; all occurrences of L).

Regarding the list of issues provided in Section 3.6, we can first observe that filling the problematic cell no. 4 is not required for calculating Attraction and Reliance scores. A second advantage is that Attraction and Reliance are straightforward descriptive measures which allow for clear and unambiguous interpretations (cf. Newman 2010: 93). Furthermore, no assumptions about the stochastic structure and the random distribution of the corpus data have to be made, since both values are simple proportions, i.e. descriptive measures.

On the downside, not taking cell no. 4 into consideration has the negative effect that the number of competing constructions, and thus the confidence one can have in the significance of the data, is not factored in. Especially in small corpora, this can have the unwelcome effect that rare nouns which happen to occur relatively frequently in the target construction produce very high reliance scores.[10] Furthermore, the fact that we use two measures instead of one can also

[10] This happened, for example, in the study by Gries et al. (2005) on the *as*-predicative construction in the case of the verbs *hail* (3 occurrences in the construction out of a total of 4; Reliance score = 75%) and *class* (3 out of 8; Reliance score = 60%), or, even more extremely, *catapult* (1 out of 1; Reliance score 100%). As discussed by the authors (Gries et al. 2005: 661–663), these scores skew the results concerning the measure of Reliance and are likely to have a strong

be seen as a disadvantage of this proposal, as a simple rank ordering is of course impossible. Attempts to conflate the values in one commonly used statistical measure such as Mutual Information or the Jaccard distance run into interpretative problems, largely caused by the very fact that the phenomenon we want to measure may in fact be two-dimensional.

4.2 Delta P (ΔP)

As has been noted, one drawback of the measures of Attraction and Reliance is that they do not take the observation concerning the number of other lexemes in other constructions rendered in cell no. 4 into account. This shortcoming is at least potentially redressed by the measure of Delta P (ΔP), which is suggested by Ellis and Ferreira-Junior (2009). As noted in Section 3.5, Ellis and Ferreira-Junior also emphasize that two reciprocal rather than one unifying measure may be required to assess the association between constructions and lexemes, on the one hand, and lexemes and constructions, on the other, and therefore recommend the use of two scores.

Technically, ΔP measures the contingent probability of a given construction attracting a given lexeme (ΔP construction → word; henceforth “ΔP Attraction”) and of a given lexeme relying on a given construction (ΔP word → construction; henceforth “ΔP Reliance”). In order to do so, it goes beyond Reliance and Attraction as such and takes into account additional information related to other probabilities. The calculation of ΔP Attraction starts out from the score for Attraction but subtracts from this score the division of cell 2 by the row total of cells 2 and 4, which relates the occurrences of other lexemes in the construction to the occurrences of all other lexemes in other constructions. Analogously, the calculation of ΔP Reliance starts out from Reliance but subtracts the division of cell 3 by the column total of cells 3 and 4. (cf. Table 7), thus taking into account the relation between the occurrences of the construction with other lexemes and the occurrences of other constructions with other lexemes.

Comparing the two ΔP scores to the Attraction and Reliance scores rendered for *suspicion* and *sign* in Table 6, one immediately notices that the corresponding scores are almost identical (if one neglects the fact that Attraction and Reliance are given as percentages). The reason for this is that the score for cell no. 4 is part of the denominator of the second division. Since this score is usually much larger

negative effect on the predictive power of Reliance (see Section 5 for more details). Effects like these can only be avoided by using larger corpora.

Table 7: Calculating scores for the two ΔP measures

$\Delta\text{P Attraction} = \frac{\text{cell 1}}{\text{cell 1} + \text{cell 3}} - \frac{\text{cell 2}}{\text{cell 2} + \text{cell 4}}$	$\Delta\text{P Attraction}_{\textit{suspicion that}}$	$= \frac{4}{548} - \frac{9}{110389}$ $= 0.0072$
	$\Delta\text{P Attraction}_{\textit{sign that}}$	$= \frac{6}{548} - \frac{54}{110389}$ $= 0.0105$
$\Delta\text{P Reliance} = \frac{\text{cell 1}}{\text{cell 1} + \text{cell 2}} - \frac{\text{cell 3}}{\text{cell 3} + \text{cell 4}}$	$\Delta\text{P Reliance}_{\textit{suspicion that}}$	$= \frac{4}{13} - \frac{544}{544 + 110830}$ $= 0.3028$
	$\Delta\text{P Reliance}_{\textit{sign that}}$	$= \frac{6}{60} - \frac{542}{542 + 110785}$ $= 0.0951$

than all the other scores in the two-by-two table, the result of the division tends to be very small, and the subtraction therefore has only a very limited effect on the result of the calculation. The same tendency can be observed for the other verbs listed in Table 5. The smaller the scores for Reliance or Attraction, the larger the relative effect of the subtraction in the ΔP scores and the resulting difference to Reliance and Attraction. Since Attraction scores tend to be lower than Reliance scores, the effects of the ΔP calculation more often lead to a re-ranking in terms of Attraction than in terms of Reliance. In general, the Attraction and Reliance approach and the ΔP approach yield very similar results, especially under the very common circumstances that the score for cell no. 4 is much higher than that for the other three cells. In many cases, therefore, the frequency adjustment achieved by the subtraction of the competing proportions in the two-by-two table has such a marginal effect that it does not seem to be worth the effort of filling cell no. 4. The less demanding measures of Attraction and Reliance may do the job just as well as the two ΔP measures.

4.3 Odds Ratio

The Odds Ratio measure (OR) provides a second way of taking the occurrences of other lexemes and other constructions into consideration. To demonstrate how it works, we again compare the nouns *suspicion* and *sign*. Recall that reliance scores are 30.769% for *suspicion* and 10.00% for *sign*; Attraction scores are 0.73% for *suspicion* and 1.09% for *sign*.

In order to explain the Odds Ratio measure, it is important to be clear about the difference between proportions, probabilities and odds. The measures of Reliance and Attraction reflect proportions in the observed data relating the frequency of the tokens of a target lexeme in a construction to either the total of the tokens of the lexeme in the corpus (Reliance), or to the total of the tokens of the construction in the corpus (Attraction). If one has sufficient trust in the dataset, one can interpret observed proportions as providing information about the probability that a certain event will happen again, in our case the probability that a given noun will occur in a construction or, more precisely, that a construction will have the values +TARGET LEXEME and +TARGET CONSTRUCTION. From the perspective of the lexeme, loosely speaking, the probability that *sign* occurs in the N+*that*-clause construction is 10% or 0.1, and that of *suspicion* 30.769% or 0.30769. The notion of odds refers to a simple transformation or function of the probability. It relates the probability which is based on what has been observed to the probability of what could also have happened given the full set of possibilities. Odds thus relate probabilities to converse probabilities. Focusing first on the variable LEXEME, the converse probability can be expressed as the number of constructions featuring the values +TARGET LEXEME and –TARGET CONSTRUCTION, or, in other words, the proportion of the tokens of the target lexemes in other constructions out of all tokens of the lexeme, yielding (60 – 6) : 60 = 0.9 for *sign* and (13 – 4) : 13 = 0.6923 for *suspicion*. The odds of the occurrence of the lexeme in the construction are then calculated by dividing the probability of their occurrence in the construction by the converse probability (cf. Table 8a). Transforming Reliance scores into fractions, these calculations can be rendered as 0.1 : 0.9 = 0.11111 for *sign* and 0.30769:0.6923 = 0.44444 for *suspicion*.[11] The resulting scores express the chance that any construction in the corpus which contains the nouns *sign* or *suspicion* respectively also contains or represents the *N+that*-clause construction. The Odds Ratio, however, as is indicated by the term *Ratio*, goes one step further than that and relates the odds of constructions containing the noun (i.e. featuring the value +TARGET LEXEME) to the odds of constructions not containing the noun, that is, the odds of the sets of constructions featuring the values –TARGET LEXEME +TARGET CONSTRUCTION VS. –TARGET LEXEME –TARGET CONSTRUCTION. This calculation is analogous to the earlier one, but it does not operate on cells nos. 1 and 2 in the contingency table, but rather on cells nos. 3 and 4 (see Table 8b). Once both odds are known, the ratio between them can be calculated by dividing the odds for constructions with the feature +LEXEME by the odds for the constructions with the feature –LEXEME (see Table 8c). The resulting Odds Ratio scores are immediately interpretable in such a way that the Odds Ratio of 90 given for *suspi-*

[11] Note that the converse probability can also be worked out by subtracting the probability (or Reliance) score from 1. For *sign* the subtraction of 0.1 from 1 yields 0.9, and for *suspicion*, the subtraction of 0.30769 from 1 yields 0.6923.

cion indicates that a construction which has the feature +TARGET LEXEME = *suspicion* is 90 times more likely to also have the feature +TARGET CONSTRUCTION = N+*that*-clause than a construction not containing *suspicion*. While we have explained the Odds Ratio in such a way as to reveal the rationale behind it, the formula in 8c can be cancelled several times yielding the simpler formula given in 8d, which produces the same results.

The calculations provided so far have depicted the situation from a reliance-based point of view, which looks at two-by-two tables from a column, i.e. vertical perspective and treats the variable LEXEME as given. The same kind of calculation can also be carried out from an attraction-based, horizontal perspective, proceeding from the variable CONSTRUCTION and beginning by calculating the odds of +TARGET CONSTRUCTION +TARGET LEXEME (cell no. 1) vs. +TARGET CONSTRUCTION – TARGET LEXEME (cell no. 3). To cut a long story short, if this were done it would

Table 8: Calculating Odds Ratio scores

a) Step 1: Calculating the odds for +TARGET LEXEME +TARGET CONSTRUCTION VS. +TARGET LEXEME –TARGET CONSTRUCTION (loosely speaking 'target lexemes in vs. outside the target construction')

$$\text{Odds target} = \frac{\text{cell 1}}{\text{cell 1 + cell 2}} : \frac{\text{cell 2}}{\text{cell 1 + cell 2}}$$

$$\text{Odds}_{\textit{suspicion that}} = \frac{4}{13} : \frac{9}{13} = 0.307601 : 0.6923 = 0.44444$$

$$\text{Odds}_{\textit{sign that}} = \frac{6}{60} : \frac{54}{60} = 0.1 : 0.9 = 0.11111$$

b) Step 2: Calculating the odds for –TARGET LEXEME +TARGET CONSTRUCTION VS. –TARGET LEXEME –TARGET CONSTRUCTION (loosely speaking 'other lexemes in vs. outside the target construction')

$$\text{Odds other} = \frac{\text{cell 3}}{\text{cell 3 + cell 4}} : \frac{\text{cell 4}}{\text{cell 3 + cell 4}}$$

$$\text{Odds}_{\textit{suspicion that}} = \frac{544}{111374} : \frac{110830}{111374} = 0.004908$$

$$\text{Odds}_{\textit{sign that}} = \frac{542}{111327} : \frac{110785}{111327} = 0.004892$$

c) Step 3: Calculating Odds Ratio by dividing odds$_{\text{target}}$ by odds$_{\text{other}}$

$$\text{Odds Ratio} = \frac{\text{odds target}}{\text{odds other}}$$

$$\text{Odds Ratio}_{\textit{suspicion that}} = \frac{0.44444}{0.004908} = 90.55$$

$$\text{Odds Ratio}_{\textit{sign that}} = \frac{0.11111}{0.004892} = 22.71$$

d) Simpler version of calculating Odds Ratio

$$\text{Odds Ratio} = \frac{\text{cell 1}}{\text{cell 2}} : \frac{\text{cell 3}}{\text{cell 4}}$$

$$\text{Odds}_{\textit{suspicion that}} = \frac{4}{9} : \frac{544}{110830} = 90.548662$$

$$\text{Odds}_{\textit{sign that}} = \frac{6}{54} : \frac{542}{110785} = 22.711199$$

eventually turn out that, for simple mathematical reasons, reliance-based and attraction-based Odds Ratio scores are in fact invariably identical. The attraction-based formula – (cell 1: cell 3):(cell 2: cell 4) – corresponding to the reliance-based one given in Table 8d would yield the same result. This mathematical property makes the use of the Odds Ratio quite attractive in our view because it can be interpreted as an overall frequency-adjusted measure for both reliance and attraction which can be used if one insists on keeping available the option of ranking lexemes on one scale in order to be able to compare them on one dimension.

Odds Ratio scores can be related to the Fisher Exact by recognizing that an Odds Ratio of 1 corresponds to the null hypothesis that there is no attraction. Small p-values of the order mentioned in earlier sections of this paper correspond to very high Odds Ratio scores, as is exemplified by the Odds Ratio score for *fact* in Table 5. However, as has been shown, identical p-values, like those rendered for *suspicion* and *sign* in Table 5, can correspond to substantially different Odds Ratio scores (22.7 as opposed to 90.5) if Reliance scores differ significantly.

Odds Ratio is superior to the Attraction and Reliance approach in that it is both frequency-adjusted and bi-directional. Like ΔP, it is superior to Fisher Exact in yielding effect sizes rather than p-values as measures of attraction, and in not relying on the stochastic nature of the data and the randomness assumption. Like ΔP and Fisher Exact, however, it does not solve the cell no. 4 problem.

4.4 Discussion

What, in summary, are the differences between the tests presented? Before we return to the list of issues provided in section 3.6, it will be illuminating to highlight quantitative similarities and differences in the outcomes of the four tests. This can be done by comparing the different rankings of the nouns as rendered by the scores. As already mentioned above, Reliance and ΔP Reliance yield identical rankings. In terms of ranks, Odds Ratio is also almost identical to Reliance, even though Odds Ratio is frequency-adjusted. ΔP Attraction and Attraction rank the collexemes in the same order, which is also very similar to that in terms of Fisher Exact. These similarities are reflected in the Spearman correlation matrix given in Table 9.

A significant but considerably weaker correlation can be observed between Attraction and especially ΔP Attraction, on the one hand, and Fisher Exact, on the other. What is remarkable here is that Odds Ratio, which is a bi-directional measure, sides with the Reliance perspective, rather than with Fisher Exact and the Attraction perspective. This is particularly interesting if we consider the following claim by Wiechmann (2008):

Table 9: Spearman correlation matrix for scores for Reliance, Attraction, Odds Ratio, Fisher Exact and the two ΔP measures

	Reliance	Attraction	Odds Ratio	Fisher Ex.	ΔP Reliance	ΔP Attraction
Reliance	1.000	0.399	0.999	−0.720	1.000	0.505
Attraction	0.399	1.000	0.411	−0.884	0.399	0.989
Odds Ratio	0.999	0.411	1.000	−0.731	0.999	0.517
Fisher Exact	−0.720	−0.884	−0.731	1.000	−0.720	−0.935
ΔP Reliance	1.000	0.399	0.999	−0.720	1.000	0.505
ΔP Attraction	0.505	0.989	0.517	−0.935	0.505	1.000

> [t]he results [. . .] give us some confidence in asserting that – should the task require it – we can go from the theoretically sound Fisher exact test, i.e. from the group of exact statistical hypothesis tests, to a maximum likelihood estimate, say odds ratios, without risking too much quantitative difference in the estimation of association strength. (Wiechmann 2008: 283)

Wiechmann's suggestion is not confirmed by our findings, since Fisher Exact correlates significantly less strongly with Odds Ratio (-0.72) than with other measures, among them Attraction, which is basically nothing other than relative frequency in the construction.

Keeping this in mind, we can now summarize how the alternative measures can cope with the problems faced by collostructional analysis. In doing so, we omit the second point in the list above, since none of the alternatives discussed here improves the situation with regard to the issue of how corpus data are sampled.

1. The randomness assumption underlying null-hypothesis testing
   This is unproblematic for the Attraction and Reliance approach, as both measures are purely descriptive. Odds Ratio also does not depend on the randomness assumption, and neither does ΔP.
2. The use of a significance measure (p-value of the Fisher Exact test) as a measure of Collostruction Strength and the resulting problems in terms of interpretability
   None of the other measures uses p-values to determine rank ordering. The scores for Attraction and Reliance, as well as those for Odds Ratio, can be directly interpreted, in the case of Odds Ratio also in terms of effect sizes. Like Attraction and Reliance, the ΔP scores do not yield effect sizes, but rather are purely descriptive statistics.

3. The effect of sample sizes on p-values
   Since none of the other measures uses p-values to measure association strength, this issue is not a problem for them.

4. The theoretical rationale behind and practical feasibility of filling cell no. 4
   The calculation of Attraction and Reliance does not require researchers to fill cell no. 4, but the price for this is that the two measures are not frequency-adjusted. In order to calculate ΔP and Odds Ratio scores the cell must be filled; thus they do not overcome this problem, which remains perhaps the most serious challenge. The frequency adjustment achieved by the ΔP scores in comparison with Attraction and Reliance seems negligible.

5. The problem of directionality – from construction to lexeme vs. lexeme to construction
   The Attraction and Reliance approach and the ΔP test yield two different measures reflecting the two directions of association. Like Collostruction Strength, Odds Ratio provides one measure, which is, however, mathematically bi-directional, yielding identical results for the Reliance and the Attraction perspectives. Interestingly, the two bi-directional measures highlight different aspects of the data: Odds Ratio is more in line with Reliance, Fisher Exact more similar to the Attraction perspective.

6. The fact that the Fisher Exact test is conditioned on the marginal distributions in the contingency table
   Since they do not test for statistical significance, none of the other four measures is subject to the same restrictions.

7. The potential effects of high absolute frequencies of lexemes outside the target construction
   The effect observed in the calculation of Fisher Exact arises from the fact that p-values are strongly determined by the marginals. A similar effect occurs for Attraction and ΔP Attraction. Reliance, ΔP Reliance and Odds Ratio do not award high ranks to lexemes with high absolute frequencies but comparatively low Reliance scores.

A candidate for an ideal measure does not emerge from this methodological discussion. The key question to be discussed in the remainder of this paper concerns the way in which and the extent of confidence with which these different measures of quantitative data retrieved from corpora can in fact be used to model lexicogrammatical associations in the minds of speakers. To discuss this question, we will first look at external experimental evidence (Section 5) and then discuss the cognitive underpinnings of representations of lexicogrammatical associations in the minds of speakers (Section 6).

5 External experimental evidence

To date, very few studies have tried explicitly to relate corpus results and measurements of lexicogrammatical associations to behavioural data elicited in a systematic and controlled way in experimental settings. Four studies stand out and will be reviewed in some detail here: Gries et al. (2005, 2010), Wiechmann (2008) and Ellis and Ferreira-Junior (2009).[12]

In a series of two papers, Gries et al. (2005, 2010) have attempted to test the psychological plausibility of collostructional analysis and to demonstrate its predictive superiority over less sophisticated ways of counting frequencies in corpora, such as relative token frequencies (cf. Bybee and Eddington 2006, and Bybee 2010: 98 for a discussion). Their test case is the *as*-predicative construction already mentioned in Section 3.3, which can be exemplified by utterances such as *He regarded him as stupid* (Gries et al. 2005: 636). As the example indicates, the construction consists of complex-transitive verbs complemented by objects and object complements that are introduced by *as*. What the authors essentially do is compare two types of indicators for the association strength between verbs and the construction derived from corpus data – collostruction strength as measured by Fisher Exact p-values and relative frequency counts of observed occurrences, i.e. basically Attraction – with the results of a sentence-completion task intended to investigate important aspects of the processing of the construction. The stimuli for the sentence-completion task were active and passive sentence beginnings of the type illustrated in 1 (Gries et al. 2005: 658). Informants were asked to provide plausible continuations.

(1) a. The biographer depicted the young philosopher
 b. The young philosopher was depicted

[12] Another relevant publication, which will be mentioned below, is Divjak (2008) on Polish *that*-constructions. Further studies that correlate behavioural data or semantic interpretations with the results of collostructional analyses but do not attempt to provide external experimental evidence, including, for example, Szmrecsanyi (2005), Gilquin (2006), Colleman (2009a and 2009b), Gries and Wulff (2009), Backus and Mos (2011) and Höche (2011), will not be discussed here.

In order to test the predictive power of the two measures of association strength vis-à-vis the behavioural data collected in the test, Gries et al. constructed four classes of verbs cross-tabulating high vs. low relative frequency and high vs. low Collostruction Strength. These classes, which are rendered in Table 10, were formed by means of the following procedure:

> we first plotted the ranks of the frequencies of all 107 verbs in the *as*-predicative against the ranks of their collostruction strength. Then, both the scalar variables Frequency and collostruction strength (CollStrength) were dichotomized into the levels high and low (disregarding the hapaxes) and combined such that we obtained four different combinations of variable levels. (Gries et al. 2005: 657)

Table 10: Classes of verbs tested by Gries et al. (2005: 657; column and row headers adapted)

	Relative frequency: high	**Relative frequency: Low**
Collostruction strength: high	*define, describe, know, recognize, regard, see, use, view*	*acknowledge, class, conceive, denounce, depict, diagnose, hail, rate*
Collostruction strength: low	*keep, leave, refer to, show*	*build, choose, claim, intend, offer, present, represent, suggest*

Corpus data and behavioural data were brought together in an analysis of variance (ANOVA) using the proportion of *as*-predicatives produced by the informants in the sentence-completion task as the dependent variable. With regard to the comparison of quantitative accounts of the data, it turned out that Collostruction Strength was a much better predictor of the experimental data than relative frequency (Attraction), which did not emerge as a significant factor at all. The three authors also tested the factor termed *Reliance* here (referred to as *faith* by them), but found that it was not significantly related to the results of the sentence-completion test either. Some interactions between potential variables were checked for significance, but the interaction between relative frequency, i.e. Attraction, and *faith* (i.e. Reliance) was not among them. The results as such are interpreted as providing evidence that Collostruction Strength is a better predictor of behaviour in a sentence completion task than Attraction.

Two remarks concerning the methods used in Gries et al.'s paper are in order. The first pertains to the classes of verbs rendered in Table 10. While these classes were formed using the method described in the quotation given above, it turns out that they subsume verbs which differ considerably with regard to their overall frequency in the corpus and their frequency of occurrence in the construction, and thus with regard to their scores for Reliance. This can be shown with reference to the top left-hand cell, which collects verbs rated high for both Collostruction Strength and relative frequency. While five of these verbs, viz. *define, describe, regard, treat* and *view,* can indeed be said to have a high to modest relative propensity to occur in the *as*-predicative, this is clearly not true for the other three verbs *know, see* and *use*. The marked difference between these two sets of verbs is revealed in Table 11, which supplies the essential quantitative information. As the table shows, *regard* stands out as the one verb that clearly specializes in occurring in the *as*-predicative construction (Reliance = 81%). *Describe, treat, view* and *define* have Reliance scores in the range of 34% to 22%, while *know, see* and *use* rate below 6%. The latter three verbs, on the other hand, boast much higher overall frequencies of occurrence in the corpus. In short, the verbs that are collected in this cell – and treated as one homogeneous group in the statistics applied to relate the quantitative data with the behavioural ones – belong to two distinct groups which should certainly not be lumped together. Over and above the frequency differences, it can also be observed that the verbs collected in this class are semantically very heterogeneous and differ grammatically with regard to the necessity of the *as*-complement, or, in fact, the necessity of their being an object complement as such. While it seems natural to consider *regard*, *treat* and *define* as trivalent verbs actually requiring the slots of subject, object and object complement to be filled (cf. ?*I regarded him,* ?*She treated me*), the verbs *see, know* and *use* are primarily divalent verbs, so that *I saw him* or *I know him* have nothing odd about them. This supports the impression created by the Reliance scores that some of these verbs, *qua* lexemes, are very strongly associated with the construction and its meaning, while others are first and foremost linked with other constructions. The method used by Gries et al. to define their classes is thus

Table 11: Results from Gries et al. 2005: 667; terminology slightly adapted

	frequency in construction	**frequency in corpus**	**Reliance/faith**	**Collostruction Strength (i.e. log-transformed p-value Fisher Exact)**
regard	80	99	80.81 %	166
describe	88	259	33.98 %	134
see	111	1,988	5.58 %	78
know	79	2,120	3.73 %	42
treat	21	92	22.83 %	28
define	18	83	21.69 %	23
use	42	1,228	3.42 %	21
view	12	41	29.27 %	17

clearly well justified, but it seems to leave potential confounding variables uncontrolled.

Secondly, there is a problem with the design of the sentence-completion test. As we have seen, collostructional analysis essentially ranks lexemes (most often verbs) according to their potential to be attracted by a construction. It thus effectively proceeds from an Attraction perspective, treating the construction as a given and the lexemes as variables.[13] The sentence-completion task exemplified in (1) above, however, takes the complementary perspective, since it uses the verbs as stimuli and asks the informants to add linguistic material that provides evidence for the activated construction. So what the tests measure strictly speaking is the behavioural counterpart to Reliance rather than to Attraction or Collostruction Strength, as Gries et al. assume.

In order to remedy at least the first problem and to use the data from the study as a test case, we have reanalyzed the data provided by Gries et al. (2005) in their appendix. Rather than dichotomizing parameters and sorting verbs into classes, we worked with the individual scores for Collostruction Strength and Reliance for all 28 verbs which were used as stimuli in the sentence-completion task. In addition, we calculated Odds Ratio scores and ΔP scores on the basis of the data available.[14] We then ranked all verbs with regard to the results from the experiment, on the one hand – awarding averaged ranks to all ties, i.e. the numerable cases where test results were identical – and the six scores, on the other (cf. Table 12), and evaluated the two-way correlations between test results and corpus-based scores by means of the Spearman rank correlation test (cf. Table 13).

While the general insights to be gained from Table 12 are summarized in Table 13, it is worth discussing selected verbs with regard to their rankings by the different measures. This is illuminating, irrespective of the relation to the experiment. Firstly, the two verbs *regard* and *describe*, which are arguably very typical of the *as*-predicative construction, are treated very similarly by all tests. Secondly, the frequent verbs *see* and *use*, which rank high in Attraction but low in Reliance, emerge with very different rankings in Reliance-based and Attraction-based measures, with Fisher Exact here clearly siding with the Attraction perspective. Thirdly, the very rare verbs *hail* and *denounce* present a problem for Reliance-

[13] As Gries' (2006) study of the meanings of the verb *run* demonstrates, this is not an inbuilt necessity. It is equally possible to turn the relation of fixed and variable elements around and investigate the constructions attracted by a given lexeme.

[14] Note that Gries et al. (2005) do not mention the reference score used to fill cell no. 4. Fortunately, one example given in Gries et al. (2010) indicates that the score was 138.565, which corresponds to the number of verb tags in ICE-GB.

Table 12: Data from Gries et al. (2005) reanalyzed in terms of ranks

Ranks Verbs	Experiment	Reliance	Attraction	Odds Ratio	Fisher Exact	ΔP Reliance	ΔP Attraction
class	1	3	15	3	10	3	12
describe	2.5	5	2	5	2	5	2
see	2.5	14	1	13	3	14	1
define	5	10	6	10	5	10	6
depict	5	8	23	8	13	8	18
offer	5	25	23	25	26	25	27
hail	7	2	15	2	9	2	11
rate	8.5	9	23	9	14	9	19
refer_to	8.5	18	11	18	16	18	10
regard	10.5	1	3	1	1	1	3
know	10.5	15	4	15	4	15	4
view	13	7	7.5	7	7	7	7
recognis|ze	13	12	7.5	12	8	12	8
denounce	13	4	15	4	11	4	13
diagnose	15	6	23	6	12	6	16
use	17.5	16	5	16	6	16	5
conceive	17.5	11	23	11	15	11	20
acknowledge	17.5	13	23	14	17	13	21
suggest	17.5	27	23	27	27	27	28
present	20.5	17	15	17	18	17	14
intend	20.5	19	23	19	21	19	22
keep	22	26	11	26	24	26	17
show	23	23	9	23	20	23	9
claim	26	20	15	20	19	20	15
choose	26	21	23	21	22	21	23
represent	26	22	23	22	23	22	24
build	26	24	23	24	25	24	26
leave	26	28	11	28	28	28	25

based measures, as the low number of observations in the construction is sufficient to give them very high Reliance scores.

With regard to the correlation with the experimental data, as shown in Table 13, Reliance, Fisher Exact and Odds Ratio correlate roughly equally well with the test results. In contrast to Gries et al.'s analysis of the data, Fisher Exact (0.667) only minimally outperforms Reliance (0.650), while Odds Ratio (0.658) also outperforms Reliance but is slightly lower than Fisher Exact. The measure of Attraction scores most poorly (0.312), while ΔP Attraction is a considerable improvement (0.506), but does not reach the other three scores. Interestingly, the correlations amongst the three leading corpus-based measures (Reliance – Fisher

Table 13: Spearman correlation matrix

	Experiment	Reliance	Attraction	Odds Ratio	ΔP Reliance	ΔP Attraction	Fisher Exact
Experiment	1.000	0.650	0.336	0.658	0.650	0.518	0.667
Reliance	0.650	1.000	0.242	0.999	1.000	0.575	0.813
Attraction	0.336	0.242	1.000	0.254	0.242	0.882	0.674
Odds Ratio	0.658	0.999	0.254	1.000	0.999	0.586	0.820
ΔP Reliance	0.650	1.000	0.242	0.999	1.000	0.575	0.813
ΔP Attraction	0.518	0.575	0.882	0.586	0.575	1.000	0.895
Fisher Exact	0.667	0.813	0.674	0.820	0.813	0.895	1.000

Exact: 0.813; Odds Ratio – Fisher Exact: 0.820; Reliance – Odds Ratio: 0.999) are considerably higher than those between the individual corpus-based measures and the experimental data. Despite considerable internal differences with regard to the actual ranking of verbs, the results of the corpus-based measures seem to be more similar to each other than to the data from the experiment. This is quite a disappointing outcome.[15] Taking further into consideration the fact that the verb *regard*, which scores the top rank in all corpus-based measures except Attraction and ΔP Attraction, is only found at position 11 in the rank list for the experimental data, we are tempted to conclude that the relation between the data from the experiment and those from the corpus is very difficult to interpret, and that this raises some doubts concerning the reliability of the experimental data as an external benchmark.

This conclusion seems to be supported by the findings from a follow-up study on the same construction (Gries et al. 2010). In this study, the authors extended the corpus basis for the quantitative part, improved the data retrieval method and used a reading-time paradigm to produce the experimental benchmark. While Collostruction Strength had only a marginally significant effect, its effect size was nevertheless twice as high as that of relative frequency, which was still not significant (Gries et al. 2010: 69–79). None of the tested interactions was found to be significant. Reliance/faith was not taken into consideration here.

A third promising attempt to relate statistical measures of associations between lexemes and constructions to experimental psychological data was made

[15] The outcome of the study by Divjak (2008), who compared the predictions of different quantitative corpus-based measures of data on the *that*-construction in Polish – among them Attraction, Reliance and Collostruction Strength – to the results of an acceptability rating, was similarly disappointing: “[I]t has beome clear that none of the measures has high predictive accuracy” (2008: 230). Interestingly, for Divjak’s dataset, which contained many combinations of verbs and constructions which were unacceptable, the measure of Reliance performed best.

by Wiechmann (2008). Wiechmann tested a wide range of association measures collected in Evert (2004), including Fisher Exact and other exact hypothesis tests, as well as likelihood measures (e.g. binomial likelihood), asymptotic hypothesis tests (e.g. z-score and t-score), point or maximum likelihood estimates of association strength (including Odds Ratio) and measures from information theory based on mutual information. Unlike Gries et al., Wiechmann did not carry out his own psycholinguistic experiments but compared the predictions of the statistical measures to the results of an eye-tracking and reading-time study published by Kennison (2001), who investigated the lexical effects of verbs on online comprehension. Specifically, Kennison looked at monotransitive verbs that preferred either NP objects (e.g. *the journalist revealed the problem . . .*) or *that*-clause objects (e.g. *the journalist admitted that the problem worried him*), but could occur in either pattern. She found that the verbs' preference for NP or *that*-clause complements was not a significant indicator of reading times. Instead, sentences with sentential complements tended to require more reading time irrespective of the preference of the verb. So the psycholinguistic benchmark used to assess the statistical tests is not very reliable, and Wiechmann (2008: 279) himself does not hesitate to mention several caveats that cast some doubt on his results having to do with statistical and behavioural uncertainties and limitations of scope. What his regression models bringing together corpus and experimental evidence indicate is that the most adequate measure is a test called *minimum sensitivity* belonging to the group of point or maximum likelihood estimates of association strength. Minimum sensitivity is a fairly simple concept, basically instructing us to compare the Reliance and Attraction scores for a given lexeme and to select the smaller of the two as a measure of the lexeme's association with the target construction. The test thus at least theoretically includes both perspectives. In addition, it has the appeal that it does not require cell no. 4 to be filled, and, like Odds Ratio, is "free from underlying distributional assumptions that are not met by natural language data" (Wiechmann 2008: 282), which remedies the problems mentioned in Section 3.1 above. However, since the test only selects one of the two measures – i.e. either Reliance or Attraction – which can usually be obtained comparatively easily, we have the impression that it loses information that is available and could be potentially important. What is more, scores for Attraction in general tend to be lower than those for Reliance, so that the ranking for Minimum Sensitivity is usually identical to that for Attraction.

Fourthly, the study by Ellis and Ferreira-Junior (2009) on constructions and their acquisition deserves further mention, although it does not supply data from experimental studies, but only corpus data of different provenance. It is in this paper that the two ΔP measures discussed in section 4.2 are introduced and applied. The authors compare several statistical tests with regard to their predictive power vis-à-vis the use of certain verbs in given constructions by second-language learners, which they take as an indicator of the mastery of a construction and the association to lexemes. The frequency of verbs in three target constructions in a learner corpus of seven non-native speakers was compared to the raw frequency of verbs in the same constructions in the output of their native speaker conversation partners. The frequency lists of these two corpora were compared to the scores produced by three measures: Collostruction Strength (Fisher Exact), ΔP Attraction and ΔP Reliance. Ellis and Ferreira-Junior interpret their findings as corroborating the hypothesis that "[t]he first-learned verbs in each construction will be those which are more distinctively associated with that construction in the input" (2009: 202). With regard to how this association is to be measured, they observe that the rank order for Collostruction Strength of native speaker data "is a very strong predictor of [non-native speaker] acquisition, as is ΔP (Construction → Word). What is less predictive is ΔP (Word → Construction)" (2009: 203). In evaluating these results one should keep in mind, however, that both the target measures and the benchmark measure are corpus-based and that both corpora are rather small and consist of contributions by a fairly limited number of speakers.

Regarding the differences between the scores arrived at, Ellis and Ferreira-Junior add an important remark concerning the two directions of associations, frequency and semantic generity:

> When a construction cues a particular word, that word occurs very often in that construction and it tends to be very generic. When a word cues a particular construction, it may be a lower frequency word, quite specific in its [. . .] semantics and thus very selective of that construction (Ellis and Ferreira-Junior 2009: 203).

This statement describes what seems to be a general pattern in many studies of lexicogrammatical association: lists of scores reflecting associations from construction to lexeme (i.e. Fisher Exact, Attraction, and ΔP Attraction) are typically headed by frequent and semantically comparatively unspecific lexemes with associations to rather wide ranges of different constructions; in the reciprocal lists of scores depicting associations from lexemes to constructions (i.e. Reliance, ΔP Reliance and also Odds Ratio), top ranks are typically occupied by less frequent or rare lexemes with fairly specific meanings. If lexemes manage to reach top ranks from both perspectives, as is the case for *fact* in the N+*that*-clause construction, we can be fairly confident that the association between lexeme and construction is indeed very strong.

6 Cognitive underpinnings

The cognitive underpinnings of the preceding considerations will be – and in fact have to be, if they are to make sense – discussed within the scope of a usage-based framework which assumes that the frequencies of repeated linguistic processing events translate into different strengths of associations in the network representing linguistic knowledge (cf. e.g. Langacker 2000; Barlow and Kemmer 2000; Bybee 2010). The more frequently a given linguistic stimulus has been processed by a speaker, the more routinized the corresponding association becomes in his or her mind. Different strengths of associations can be understood as representing different degrees of *entrenchment* in the network. More deeply entrenched associations are reflected behaviourally in higher degrees of routinization and automatization and lower levels of cognitive effort required for processing (cf. Schmid 2007, forthcoming and Blumenthal-Dramé 2012 for more details).[16]

While all this seems to be fairly straightforward and in line with the well-attested frequency effect in language and elsewhere (cf. e.g. Diessel 2007; Knobel et al. 2008), the complexity of the multifarious ways in which repeated exposure to linguistic structures affects the network has arguably been underestimated so far (cf. Schmid 2010). The main reason for this is that the network consists of different types of associations that constantly and simultaneously conspire and compete with each other to set up connections and affect degrees of entrenchment. Therefore, if one wishes to understand the ways in which frequency of processing may have an effect on lexicogrammatical attraction phenomena, it is crucial to differentiate these associations and to reflect on their interactions.

To begin with, the communicative potential of language is based on *symbolic associations* (Saussure 1916: 98) which connect the forms and meanings of signs and constructions of different types of complexity. These range from individual words or even morphemes (*table, come, un-*), to compositional and non-compositional lexically-filled patterns (*join the army, shoot the breeze*), partly-filled schemas (*the X-er the Y-er*; Fillmore et al. 1988: 506) and fully schematic argument-structure constructions (e.g. the ditransitive, resultative, caused-motion construction; Goldberg 1995). Looking at symbolic associations in isolation, the picture seems to be as simple as explained above: the more frequently we associate the form of a word or construction with a given meaning, the

[16] Blumenthal-Dramé (2012) reports on studies applying a range of psycholinguistic and neurolinguistic tests to measuring entrenchment. This work is not reviewed in greater detail here as it deals with entrenchment on the level of derivational morphology rather than the lexicogrammatical level.

stronger the symbolic association between the form and the meaning. Upon seeing or hearing the form, we immediately and effortlessly activate the meaning in comprehension, and vice versa in production. This frequency effect treats lexical items and constructions, somewhat naively, out of cotext and context, as if they occurred in isolation. The corresponding type of entrenchment can therefore be referred to as *cotext-free entrenchment* (Schmid 2010: 120). If one is willing to accept corpus data and the data in two-by-two contingency tables as proxies for frequency of occurrence in language use,[17] the total number of occurrences of a target lexeme – i.e. the column total of the first column – can be regarded as a rough indicator of its degree of cotext-free entrenchment, and the total number of occurrences of the target construction – i.e. the row total of the upper row – as a rough indicator of the cotext-free entrenchment of the construction.

As hinted at in the previous paragraph, cotext-free entrenchment should in fact rather not be seen as being directly reflected in absolute frequencies in the corpus as such. This warning is important. The degree of entrenchment of a lexeme or construction is inevitably linked to degrees of entrenchment of other lexemes or constructions that potentially compete for the encoding of a given idea. Even cotext-free entrenchment is never entrenchment as such, but has an inbuilt paradigmatic component and rests on a second type of association in the network: *paradigmatic associations*. From the speaker's perspective, given a certain idea to be encoded, cotext-free entrenchment captures the likelihood with which a given lexeme or construction is selected in tacit and unconscious comparison with other lexemes and constructions that are potentially activated by the target idea as well**.** This means that cotext-free entrenchment must be operationalized from an onomasiological point of view (Geeraerts 2006: 85) as "conceptual frequency" (Hoffmann 2005: 107–110) or "paradigmatic relative frequency" (Schmid forthcoming), taking into account "paradigmatic salience" (Blumenthal-Dramé 2012: 191). These notions essentially capture the cell no. 4

[17] Note that the link between corpus data, which represent frequency of usage and degrees of entrenchment, is much less direct than is often suggested (cf. Blumenthal-Dramé 2012: 28–33, 44–65). On the one hand, corpus data can do no more than serve as a proxy of frequency of use in the speech community or the sections thereof captured in the corpus. Entrenchment, on the other hand, is an individual cognitive phenomenon. Interpreting corpus data as evidence for degrees of entrenchment thus means building a rather shaky bridge across the gap between the usage of, ideally, thousands of different speakers whose output is collected in a corpus, and the cognitive system of an idealized speaker-hearer not at all unlike the one envisaged by Chomsky. Recent studies on individual differences by Dąbrowska and others (cf. e.g. Dąbrowska 2008; Street and Dąbrowska 2010) have begun to explore the gap between conventionalized grammar and individual entrenchment.

problem discussed in Section 3.4. The score to be inserted in this cell is defined as reflecting the number of all other lexemes in all other constructions. As was emphasized in Section 3.4, this should be interpreted as involving the paradigmatic competitors of the target lexeme and the target construction.

Symbolic associations thus inevitably interact with paradigmatic associations in bringing about different degrees of cotext-free entrenchment. In order to quantify the cotext-free entrenchment of symbolic associations, the total numbers of occurrences of the target lexeme and the target construction have to be related to the total number of occurrences of paradigmatically competing lexemes and constructions. As indicated by the dashed arrows in Figure 1 below, to achieve this using the scores available in two-by-two contingency tables, the column total of the first column and the row total of the first row have to be related to the score in cell no. 4.

When it comes to measuring lexicogrammatical attraction phenomena, a third type of association in the network, *syntagmatic associations*, must be taken into consideration in addition to symbolic and paradigmatic associations. Syntagmatic associations link symbolic associations to linguistic forms and meanings which co-occur in linear sequence in usage. They are invariably activated in all linguistic processing events, as the computation of sentence and discourse meanings requires the binding of sequentially presented information. For example, clause subjects have to be brought together grammatically and semantically with predicates, modifiers with heads, and so on. The combined processing creates syntagmatic associations between the symbolic associations activated by the individual elements. More importantly, especially in the present context, repeated identical or similar syntagmatic associations routinize in such a way that they become the cognitive substrate of syntagmatic attraction phenomena such as collocations and collostructions. One part of a collocation, say *commit*, can prime (cf. Hoey 2005) the activation of other parts of the larger structure (e.g. *crime, murder, suicide*, etc.) as a result of the routinized syntagmatic association resulting from prior repeated co-processing. Likewise, one part of a collostruction has the potential to cue the activation of the other part. Because of the linear nature of language, linguistic elements thus never cue only symbolic associations linking forms and meanings, but invariably also syntagmatic associations linking them to other linguistic elements and constructions frequently co-occurring in the immediate cotext. A good example of how symbolic and syntagmatic associations interact and even compete is the comprehension of non-compositional chunks such as idioms, in which the symbolic associations activated by individual forms (e.g. *shoot* ‘shoot’ and *breeze* ‘breeze’) are largely inhibited by the syntagmatic associations between the elements in the chunk (*shoot* → *breeze*), which results in the activation of a second-order symbolic association connecting the whole chunk *shoot the breeze* to the idiomatic meaning 'engage in casual talk'.[18] The cotext-free entrenchment of the lexemes *shoot* and *breeze* is overruled here by the strong *cotextual entrenchment* of the syntagmatic association between the two words and the cotext-free entrenchment of the fixed phrase (cf. Schmid 2010: 120–125, and Schmid forthcoming for more details).

In the two-by-two contingency table, hypothetical strengths of syntagmatic associations are first and foremost reflected in the scores for Attraction and Reliance, i.e. the relations between cell no. 1 (lexeme in construction) and all occurrences of the lexeme (column total of first column) and the relation between cell no. 1 and all occurrences of the construction (row total of first row). In Figure 1 below, these relations are indicated by dotted arrows and boxes. Crucially, however, paradigmatic competitors have to be taken into account with regard to cotextual entrenchment as well, since the target construction potentially activates other, paradigmatically related lexemes, and the target lexeme activates other constructions paradigmatically related to the target construction. Of course, the overall amount of these paradigmatic competitors is factored in by relating the occurrences of the lexeme in the construction (cell no. 1) to the occurrences of other lexemes in the construction (cell no. 3), on the one hand, and to the occurrences of the lexeme in other constructions (cell no. 2), on the other. But these numbers arguably do not tell the whole story. The extra effect we see at work here can be explained with an analogy to the stock market: if only one single powerful investor holds, say, 45% of the shares of a company in his portfolio, while the rest of the shares are scattered among thousands of private shareholders, their weight in the company will be stronger than if there are other bigger players controlling, for example, 35% or 30% of the shares. The significance of one portfolio is thus not only determined by the simple proportion of shares, but also by both the number of other shareholders and the sizes of their portfolios. Likewise, the significance of the association of a given lexeme to a given construction depends not only on its Reliance score, but also on the number of other constructions in which it can occur and on the lexeme's Reliance scores to these. In short, the *dispersion* profile of the lexeme plays a role. If a lexeme occurs reasonably frequently in a given construction but does not rely on it to a great extent – as was the case, for instance, for the noun *sign* in the N+*that*-clause construction or the verb *see* in the *as*-predicative construction – then the lexeme's strength of association to this construction will depend on whether its other uses are consolidated or strongly dispersed, in particular on whether or not it relies on some other constructions

[18] Cf. Blumenthal-Dramé (2012: 76–80) on gestalt-psychological foundations of chunking and the so-called "top-down coercion" of chunks to their parts.

for a larger proportion of its occurrences. Of course, neither of these pieces of information is contained in the score entered in cell no. 3. Couched in terms of complex adaptive systems theory (cf. e.g. Ellis and Larsen-Freeman 2009), the question is whether the target construction is the only or major *attractor* for a lexeme, or whether there are other nodes in the network that strongly invite the lexeme to enter into a syntagmatic association. The same line of reasoning applies to cell no. 2, which should also contain information about the number of competing lexemes and their Attraction scores. If a given lexeme competes with a small number of other collexemes, the significance of its Attraction score certainly differs from a situation where there is a large range of collexemes. In this case, the situation is slightly less complex, since at least the number of competing lexemes is usually known and remains constant for all lexemes targeted. Nevertheless, it is important to know, among other things, whether or not there are stronger competitors around. In sum, the scores in cells nos. 2 and 3 representing the occurrences of the target lexeme in other constructions and of other lexemes in the target construction do not do justice to additional paradigmatic associations which also define the place of a node in the network. These would have to be looked at much

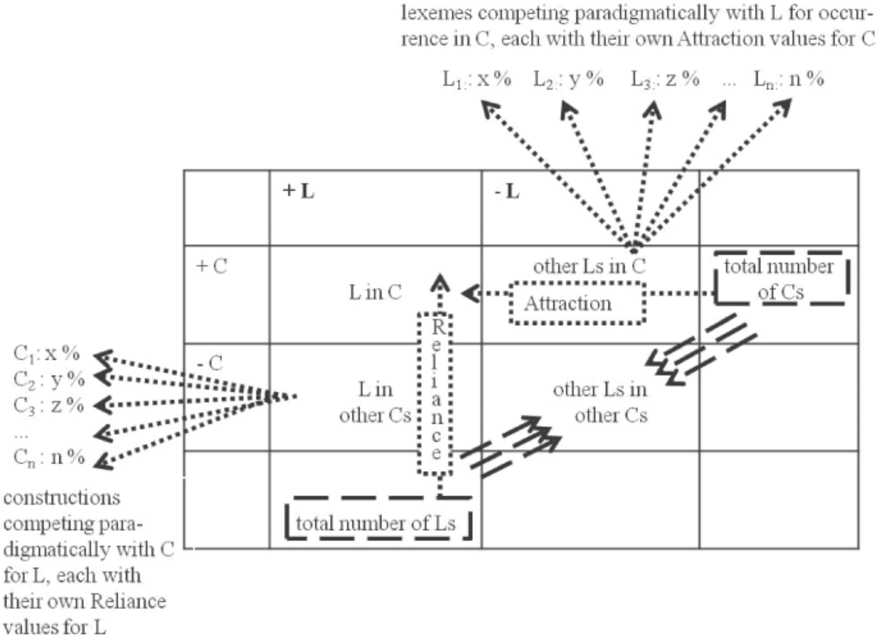
lexemes competing paradigmatically with L for occurrence in C, each with their own Attraction values for C
L1: x %
L2: y %
L3: z %
...
Ln: n %
+ L
- L
+ C
L in C
other Ls in C
Attraction
total number of Cs
- C
L in other Cs
Reliance
other Ls in other Cs
total number of Ls
C1: x %
C2: y %
C3: z %
...
Cn: n %
constructions competing paradigmatically with C for L, each with their own Reliance values for L

Fig. 1: Cotext-free and cotextual entrenchment as reflected in contingency tables (dashed arrows and boxes indicate scores and relations influencing cotext-free entrenchment; dotted arrows and boxes indicate scores and relations influencing cotextual entrenchment)

more closely with the aim of factoring in the precise distributional properties underlying these scores, including the numbers and Reliance scores of other constructions in which the lexeme occurs and the numbers and Attraction scores of other lexemes that occur in the construction. This is represented in Figure 1 by means of the dotted arrows pointing outwards from cells nos. 2 and 3.

From a cognitive perspective, then, hypothetical strengths of lexicogrammatical attraction phenomena are determined by a very complex interaction of different strengths of the symbolic associations of the lexemes and constructions involved, of the syntagmatic associations between lexemes and constructions, and of no less than four types of paradigmatic associations. The interaction of these associations produces a complex interplay of cotext-free and cotextual entrenchment which so far has not been understood in any detail. What these interactions mean in terms of the data represented in two-by-two contingency tables, and to what extent they go beyond the scope of these datasets, is summarized in Figure 1. Here, the scores and relations influencing cotext-free entrenchment are indicated by dashed lines, and the scores and relations influencing cotextual entrenchment by dotted ones.

Hints to some of these complications can be found in the literature.[19] In identifying the meaning of the most dominant lexical anchor of the *as*-predicative construction, Gries et al. (2005: 652–654) draw on the degree to which a target lexeme relies on other constructions as an argument. They observe that the verbs *see* and *consider* are more strongly associated with other constructions and therefore do not qualify as well as *regard* does. *See* is attracted by the monotransitive and other constructions, and *consider* mainly by the complex-transitive construction without *as*, cf. their example *He considered his marriage indissoluble* (2005: 637). While the authors are thus clearly aware of the challenge theoretically, their claim that collostructional analysis can handle the problem is only plausible for *consider* but certainly not for *see*, which ranks third with regard to collostruction strength because of its high Attraction value and its high overall frequency in the corpus. This rank does not reflect the fact that the verb *see* is associated with other attractor constructions that presumably exert stronger forces than the *as*-predicative. The paradigmatic information that would be required here does not come forward as some kind of automatic by-product of calculating Collostruction Strength. It is perhaps more revealing here to include the Reliance perspective, as both verbs indeed score low results according to the data provided by Gries et al., with Reliance scores of 5.58% for *see* and 3.41% for *consider*, as com-

[19] Cf. Blumenthal-Dramé (2012: 191 *et passim*) on paradigmatic aspects relevant for the entrenchment of suffix-derivations.

pared to 80.01% for *regard*. What must be added, however, is that none of the measures discussed in the present paper alone does an adequate job of taking into account other constructions which strongly attract a target lexeme.

Arguably, it is precisely the complexity sketched out in this section that has constituted the major challenge for the statistical tests discussed in this paper.[20] The Reliance and Attraction approach falls short of taking cotext-free entrenchment into account, since it does not take cell no. 4 into consideration and thus misses out on integrating cotext-free entrenchment and its paradigmatic aspects. In principle, the two ΔP scores do take this information into consideration, but the mathematical effects of this extra step and the theoretical gain are very limited. While scores for ΔP Attraction can indeed diverge from Attraction and improve results, as we have seen in Section 5, those for Reliance and ΔP Reliance tend to be almost identical. Fisher Exact does integrate information on cotext-free entrenchment, but seems, at least in comparison with Odds Ratio, to give far too much weight to it while strongly neglecting key aspects of cotextual entrenchment. As a result, lexemes which presumably reach high levels of cotext-free entrenchment because of their high frequency of occurrence tend to emerge as being strongly associated with constructions, even though their uses in the construction do not account for a large proportion of their uses in the corpus, and even though they are more strongly associated with other constructions. As shown in Section 3.5, Fisher Exact also suggests that the strength of syntagmatic associations can be the same for combinations of high cotext-free and low cotextual entrenchment (the scenario in Table 4b), on the one hand, and for combinations of low cotext-free and high cotextual entrenchment (the scenario in Table 4a), on the other. Odds Ratio takes cotext-free entrenchment into account, but propels rarer lexemes with hypothetically lower degrees of cotext-free entrenchment (e.g. *suspicion*) into fairly high ranks. Whether either of these two tests, which, as we have seen, produce rather divergent results, indeed manages to capture the mutual attraction of lexemes and construction in one measure must presumably remain an open question for as long as we remain unable to produce adequate external evidence to properly assess their predictive power. This conclusion would suggest that, as argued by Ellis and Ferreira-Junior (2009) and Schmid (2010), it might be wise to keep the two directions of association apart and work with two reciprocal measures in the meantime. What has become clear is that none of the measures manages to include information on the paradigmatic di-

[20] As discussed in Schmid (forthcoming) the interaction of entrenchment, frequency and salience is even more complex as, in addition to the types of associations and salience discussed in the text here, pragmatic associations and their salience, i.e. pragmatic salience, have to be taken into consideration.

mension of cotextual entrenchment represented by the dotted arrows pointing outside the contingency table in Figure 1, since all of them are unable to integrate data on the numbers and Reliance scores of paradigmatically competing constructions and the numbers and Attraction scores of paradigmatically competing lexemes.

7 Conclusion and outlook

In light of the considerable imponderables that have come to the surface in the previous section, the critical discussion of collostructional analysis and other measures of lexicogrammatical associations presented earlier in this paper may seem rather petty, even insignificant perhaps. What, one may well ask, is the point of engaging in quibbles about details of statistical tests and tools if the cognitive underpinnings of what their results are supposed to model or predict are far from clear anyway? Our discussion of cognitive underpinnings of lexicogrammatical associations suggests that at present we do not seem to know enough about how frequency of usage affects different types of associations and determines the entrenchment of syntagmatic associations between constructions and lexemes in order to adequately assess the measures discussed in this paper from an Archimedean, i.e. noncorpus-based, vantage point. Wiechmann's caveat that "there is still a strong need for empirical evaluations of competing measures of collocativity (or collostruction strength for that matter)" (2008: 283) remains as valid today as it was in 2008. Linguists are presumably well advised to refrain from being overly enthusiastic about their capability to actually model cognitive phenomena with the help of quantitative evidence gleaned from corpora.

While it has become obvious that the path pioneered by Gries et al. (2005, 2010), Divjak (2008) and Wiechmann (2008) – as well as Blumenthal-Dramé (2012) in the field of derivational morphology – in their search for converging evidence is a very promising one, experimental methods have to be improved considerably before we can have more confidence in assessing the validity of research findings based on corpus data. We hope that our paper has demonstrated what different corpus-based measures can do, but also what they are so far unable to do. To conclude with concrete recommendations, if one insists on producing a single rank order of collexemes for a given construction, then one should be aware that Fisher Exact mainly looks at the association from the perspective of constructions to lexemes, tends to highlight the role of frequent collexemes at the cost of specific ones, and does not provide an effect size measure. Odds Ratio, which also produces one single rank order, manages to produce effect sizes, but emphasizes the Reliance of lexemes on constructions. Both measures require a score for cell no. 4, but this comes with the advantage that they take competing probabilities into account. If one decides that the cognitive implications are not yet sufficiently well understood to warrant the use of a single measure, then we recommend working either directly with Attraction and Reliance scores or with the two reciprocal Delta P measures, which take sample size into consideration but of course also have the concomitant drawback that cell no. 4 must be filled.

Acknowledgements

We would like to thank Susanne Grandmontagne, Franziska Günther, Daphné Kerremans and three anonymous reviewers for their invaluable comments on an earlier version of this paper and their extremely helpful suggestions.

References

Baayen, Harald. 2011. Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics* 11, 295–328.

Backus, Ad & Maria Mos. 2011. Islands of (im)productivity in corpus data and acceptability judgements: Contrasting two potentiality constructions in Dutch. In D. Schönefeld (ed.), *Converging evidence*, 165–192, Amsterdam: Benjamins.

Barlow, Michael & Suzanne Kemmer (eds.). 2000. *Usage-Based Models of Language*. Stanford, CA: CSLI Publications.

Blumenthal-Dramé, Alice. 2012. *Entrenchment in usage-based theories. What corpus data do and do not reveal about the mind*. Berlin: Walter de Gruyter.

Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge/New York: Cambridge University Press.

Bybee, Joan & David Eddington. 2006. A usage-based approach to Spanish verbs of 'becoming'. *Language* 82(2), 323–355.

Church, Kenneth W. & Patrick Hanks. 1990. Word association norms, mutual information & lexicography. *Computational Linguistics* 16(1), 22–29.

Clear, Jeremy. 1993. From Firth principles. Computational tools for the study of collocation. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.), *Text and technology. In honour of John Sinclair*, 271–292. Philadelphia/Amsterdam: Benjamins.

Colleman, Timothy. 2009a. Verb disposition in argument structure alternations: A corpus study of the Dutch dative alternation. *Language Sciences* 31, 593–611.

Colleman, Timothy. 2009b. The semantic range of the Dutch double object construction. A collostructional perspective. *Constructions and Frames* 1, 190–220.

Colleman, Timothy. 2010. Beyond the dative alternation: The semantics of Dutch *aan*-Dative. In D. Glynn & K. Fischer (eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches*, 271–303. Berlin: Mouton de Gruyter.

Dąbrowska, Ewa. 2008. The later development of an early-emerging system: The curious case of the Polish genitive. *Linguistics* 46, 629–650.

Diessel, Holger. 2007. Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology* 25, 108–127.

Divjak, Dagmar. 2008. On (in)frequency and (un)acceptability. In Barbara Lewandowska-Tomaszczyk (ed.), *Corpus linguistics, computer tools and applications – state of the art*, 213–233. Frankfurt: Peter Lang.

Ellis, Nick C. & Diane Larsen-Freeman (eds.). 2009. *Language as a Complex Adaptive System*. Chichester: Wiley-Blackwell.

Ellis, Nick & Fernando Ferreira-Junior. 2009. Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal* 93(3), 370–385.

Evert, Stefan. 2004. The statistics of word cooccurrences: Word pairs and collocations. Unpublished doctoral dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Evert, Stefan & Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, *Toulouse, France*, 188–195.

Fillmore, Charles J., Paul Kay, & M. Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64, 501–538.

Fries, Charles C. 1952. *The structure of English*. London: Longmans, Green and Company.

Geeraerts, Dirk. 2006. *Words and other wonders: Papers on lexical and semantic topics*. Berlin/New York: Walter de Gruyter.

Gilquin, Gaëtanelle. 2006. The verb slot in causative constructions. Finding the best fit, *Constructions*. http://elanguage.net/journals/index.php/constructions/article/view/18/23 [Accessed March 2013].

Glynn, Dylan. 2010. Corpus-driven cognitive semantics. Introduction to the field. In D. Glynn & K. Fischer (eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches*, 1–41. Berlin etc.: de Gruyter Mouton.

Goldberg, Adele E. 1995. *Constructions. A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goodman, Steven N. 2002. P value. In R. C. Elston, J. M. Olson & L. Palmer, (eds.), *Biostatistical genetics and genetic epidemiology*, 591–594. Chichester: John Wiley and Sons.

Gries, Stefan Th. 2005. Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory* 1(2), 277–94.

Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: The many meanings of *to run*. In S. Th. Gries & A. Stefanowitsch (eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*, 57–99. Berlin/New York: Mouton de Gruyter.

Gries, Stefan Th. 2011. Corpus data in usage-based linguistics: What's the right degree of granularity for the analysis of argument structure constructions? In Mario Brdar, Stefan Th. Gries & Milena Žic Fuchs (eds.), *Cognitive linguistics: Convergence and expansion*, 237–256. Amsterdam/Philadelphia: John Benjamins.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9, 97–129.

Gries, Stefan Th. & Anatol Stefanowitsch. 2010. Cluster analysis and the identification of collexeme classes. In J. Newman & S. Rice (eds.), *Experimental and empirical methods in the study of conceptual structure, discourse, and language*, 73–90. Stanford: CSLI.

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 J. Newman & S. Rice (eds.), *Experimental and empirical methods in the study of conceptual structure, discourse, and language* (eds.), 59–72. Stanford: CSLI.

Gries, Stefan Th. & Stefanie Wulff. 2009. Psycholinguistic and corpus linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics* 7, 163–186.

Hampe, Beate. 2011. Discovering constructions by means of collostruction analysis: The English denominative construction. *Cognitive Linguistics* 22(2), 211–245.

Hilpert, Martin. 2006. Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory* 2(2), 243–256.

Hilpert, Martin. 2010. The force dynamics of English complement clauses: A collostructional analysis. In D. Glynn & K. Fischer (eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches*, 155–178. Berlin: Mouton de Gruyter.

Höche, Silke. 2011. *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*, 115–142. Amsterdam: Benjamins.

Hoey, Michael. 2005. *Lexical priming. A new theory of words and language.* London/New York: Routledge/Taylor & Francis.

Hoffmann, Sebastian. 2005. *Grammaticalization and English complex prepositions: A corpus-based study*. New York: Routledge.

Hunston, Susan & Gill Francis. 2000. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*, Amsterdam/Philadelphia: Benjamins.

Kay, Paul & Charles J. Fillmore. 1999. Grammatical constructions and linguistic generalizations: The What's X doing Y? construction. *Language* 75, 1–33.

Kennison, Sheila M. 2001. Limitations on the use of verb information during sentence comprehension. *Psychonomic Bulletin & Review* 8(1), 132–138.

Kilgarriff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1/2, 263–276.

Knobel, Mark, Matthew Finkbeiner & Alfonso Caramazza. 2008. The many places of frequency: Evidence for a novel locus of the lexical frequency effect in word production. *Cognitive Neuropsychology* 25(2), 256–286.

Langacker, Ronald W. 2000. A dynamic usage-based model. In M. Barlow & S. Kemmer (eds.), *Usage-based models of language*, 1–63. Stanford: CSLI Publications.

Lipka, Leonhard. 2002. *An outline of English lexicology. Lexical structure, word semantics & word-formation*. Tübingen: Narr.

Manning, Chris, & Hinrich Schütze. 2001. *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Mukherjee, Joybrato & Stefan Th. Gries. 2009. Collostructional nativisation in New Englishes. Verb-construction associations in the International Corpus of English. *English World-Wide* 30(1), 27–51.

Newman, John. 2010. Balancing acts: Empirical pursuits in Cognitive Linguistics. Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches*, 79–99. Berlin: de Gruyter Mouton.

Owen Frank & Ronald Jones. 1977 *Statistics*. Stockport: Polytech Publishers.

Saussure, Ferdinand de. 1916. *Cours de linguistique générale*. Paris: Payot.

Schmid, Hans-Jörg. 2000. *English abstract nouns as conceptual shells. From corpus to cognition*, Berlin/New York: Mouton de Gruyter.

Schmid, Hans-Jörg. 2007. Non-compositionality and emergent meaning of lexico-grammatical chunks: A corpus study of noun phrases with sentential complements as constructions. *Zeitschrift für Anglistik und Amerikanistik*, 3(3), 313–340.

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

Schmid, Hans-Jörg. forthcoming. Lexico-grammatical patterns, pragmatic associations and discourse frequency. In T. Herbst, H.-J. Schmid & S. Faulhaber (eds.), *Constructions – collocations – patterns*. Berlin: Mouton de Gruyter.

Sinclair, John M. 1991. *Corpus, concordance, collocation*, Oxford: Oxford University Press.

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

Stefanowitsch, Anatol & Stefan Th. Gries. 2008. Channel and constructional meaning: A collostructional case study. In G. Kristiansen & R. Dirven (eds.), *Cognitive sociolinguistics*, 129–152. Berlin/New York: Mouton de Gruyter.

Stefanowitsch, Anatol & Stefan Th. Gries. 2009. Corpora and grammar. In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics: An international handbook, Volume 2*, 933–951. Berlin/New York: Mouton de Gruyter.

Street, James & Ewa Dąbrowska. 2010. More individual differences in Language Attainment: How much do adult native speakers of English know about passives and quantifiers? *Lingua* 120, 2080–2094.

Stubbs, Michael. 1995. Collocations and semantic profiles. On the cause of the trouble with quantitative studies. *Functions of Language* 2, 23–55.

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), 113–150.

Wiechmann, Daniel. 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.

Wulff, Stefanie. 2005. Online statistics labs. *Corpus Linguistics and Linguistic Theory* 1(2), 303–308.

Zeschel, Arne. 2009. What's (in) a construction? In V. Evans & S. Pourcel (eds.), *New Directions in Cognitive Linguistics*, 185–200. Amsterdam/Philadelphia: Benjamins.