Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics

Some necessary clarifications*

Stefan Th. Gries
University of California, Santa Barbara

In the last few years, a particular quantitative approach to the syntax-lexis interface has been developed: collostructional analysis (CA). This approach is an application of association measures to co-occurrence data from corpora, from a usage-based/cognitive-linguistic perspective. In spite of some popularity, this approach has come under criticism in Bybee (2010), who criticizes the method for several perceived shortcomings and advocates the use of raw frequencies/percentages instead. This paper has two main objectives. The first is to refute Bybee's criticism on theoretical and empirical grounds; the second and further-reaching one is to outline, on the basis of what frequency data *really* look like, a cline of analytical approaches and, ultimately, a new perspective on the notion of construction based on this cline.

1. Introduction

Linguistics is a fundamentally divided discipline, as far as theoretical foundations and empirical methodology are concerned. On the one hand and with some simplification, there is the field of generative grammar with its assumptions of (i) a highly modular linguistic system within a highly modular cognitive system (ii) with considerable innate structure given the poverty of the stimulus, and (iii) a methodology largely based on made-up judgments of made-up (often context-free) sentences. On the other hand and with just as much simplification, there is the field of cognitive/functional linguistics with its emphasis on (i) domain-general mechanisms, (ii) pattern-learning based on statistical properties of the input, and (iii) an (increasing) reliance on various sorts of both experimental and observational data.

Over the last 25+ years, this latter field has amassed evidence calling into the question the assumption of a highly modular linguistic system, a large amount of

Studies in Language 11:3 (2012), 477–510. DOI 10.1075/sl.11.3.02gri
ISSN 0378–4177 / E-ISSN 1569–9978

innate structure, and the reliability of the predominant kind of acceptability judgment data. First, there is now a lot of experimental evidence that shows how much aspects of syntax interact with, or are responsive to, e.g., phonology, semantics, or non-linguistic cognition. Second, many studies have now demonstrated that the supposedly poor input is rich in probabilistic structure, which makes many of the supposedly unlearnable things very learnable. Third, Labov and Levelt, among others, already showed in the early 1970s that the judgments that were adduced to support theoretical developments were far from uncontroversial and that better ways of gathering judgment data are desirable. Over the last few years, corpus data have especially become one of the most frequently used alternative types of data.

This movement towards empirically more robust data is desirable. However, while (psycho)linguistic experimentation has a long history of methodological development and refinement, the situation is different for corpus data. While corpus linguistic approaches have been around for quite a while, the methodological evolution of corpus linguistics is still a relatively young development and many corpus-based studies are lacking the methodological sophistication of much of the experimental literature. This situation poses a bit of a challenge because, while a usage-based approach to language — an approach stipulating that the use of language affects the representation and processing language — does not *require* usage data, the two are of course highly compatible. This makes the development of an appropriate corpus-linguistic toolbox an important goal for usage-based linguistics.

This paper is concerned with a recent corpus-based approach to the syntax-lexis interface called collostructional analysis (CA), which was developed to apply recent developments in corpus linguistics to issues and questions in cognitive/usage-based linguistics. Most recently, however, this approach was criticized (Bybee 2010: Section 5.12) for several perceived shortcomings. The first part of this paper constitutes a response to Bybee's claims, which result from a lack of recognition of the method's assumptions, goals, and published results. However, I will also discuss a variety of cognitive-linguistic and psycholinguistic notions which are of relevance to a much larger audience than just collostructional researchers and which speak to the relation between data and the theory supported or required by such data. Section 2 provides a brief explanation of the collostructional approach — while the approach is now reasonably widespread, this is necessary for the subsequent discussion. Section 3 presents the main claims made by Bybee, which I will then address in Section 4. Section 5 will develop a cline of co-occurrence complexity and discuss its theoretical motivations and implications with a variety of connections to psychological and psycholinguistic work.

2. Collostructional analysis: A brief overview

2.1 Perspective 1: CA and its goals

All of corpus linguistics is by definition based on frequencies — either on the question of whether something occurs (i.e., is a frequency $n>0$?) or not (i.e., is $n=0$?) or on the question of how often something occurs (how large is n?) — which makes it a distributional discipline. Since linguists are usually not that much interested in frequencies *per se* but rather structure, semantics/meaning, pragmatics/function, etc., corpus-linguistic work has to make one very fundamental assumption, namely that distributional characteristics of an element reveal many if not most of its structural, semantic, and pragmatic characteristics; cf. the following quote by Harris (1970: 785f.):

> [i]f we consider words or morphemes *A* and *B* to be more different in meaning than *A* and *C*, then we will often find that the distributions of *A* and *B* are more different than the distributions of *A* and *C*. In other words, difference of meaning correlates with difference of distribution.

A more widely-used quote to make the same point is Firth's (1957: 11) "[y]ou shall know a word by the company it keeps." Thus, corpus-linguistic studies of words have explored the elements with which, say, words in question co-occur, i.e., the lexical items and, to a much lesser degree, grammatical patterns with which words co-occur — their *collocations* and their *colligations*. However, since some words' overall frequencies in corpora are so high that they are frequent nearly everywhere (e.g., function words), corpus linguists have developed measures that downgrade/penalize words whose high frequency around a word of interest *w* may reflect more their overall high frequency than their revealing association with *w*. Such measures are usually referred to as *association measures* (AMs) and are usually applied such that one

i. retrieves all instances of a word *w*;
ii. computes an AM score for every collocate of *w* (cf. Wiechmann 2008 or Pecina 2009 for overviews);
iii. ranks the collocates of *w* by that score;
iv. explores the top *t* collocates for functional patterns (where *functional* encompasses 'semantic', 'pragmatic', 'information-structural', …).

Thus, the purpose of ranking words on the basis of such AMs is to produce a ranking that will place words at the top of the list that (i) have a relatively high frequency around *w* while (ii) not being too frequent/promiscuous around other words.

2.2 Perspective 2: CA and its mathematics/computation

CA is the extension of AMs from lexical co-occurrence — a word w and its lexical collocates — to lexico-syntactic co-occurrence: a construction c and the x words w_1, w_2, …, w_x in a particular slot of c. Thus, like most AMs, CA is based on (usually) 2×2 tables of observed (co-)occurrence frequencies such as Table 1.

Table 1. Schematic frequency table of two elements *A* and *B* and their co-occurrence

	B	$\neg B$	**Totals**
A	$nA \;\&\; B$	$nA \;\&\; \neg B$	nA
$\neg A$	$n\neg A \;\&\; B$	$n\neg A \;\&\; \neg B$	$n\neg A$
Totals	nB	$n\neg B$	$nA \;\&\; B \;\&\; \neg A \;\&\; \neg B$

Two main methods are distinguished. In the first, *collexeme analysis* (cf. Stefanowitsch & Gries 2003), A is a construction (e.g., the ditransitive NP V NP1 NP2), $\neg A$ corresponds to all other constructions in the corpus (ideally on the same level of specificity), B is a word (e.g., *give*) occurring in a syntactically-defined slot of such constructions, and $\neg B$< corresponds to all other words in that slot in the corpus. A collexeme analysis requires such a table for all x different types of B in the relevant slot of A. For example, Table 2 shows the frequency table of *give* and the ditransitive based on data from the ICE-GB. Each of these x tables is analyzed with an AM; as Stefanowitsch & Gries (2003: 217) point out, "[i]n principle, any of the measures proposed could be applied in the context of CA." Most applications of CA use the p-value of the Fisher-Yates exact test (p_{FYE}) or, as a more easily interpretable alternative, the (usually) negative $\log_{10}$ of that p-value (cf. Gries, Hampe & Schönefeld 2005: 671f., n. 13).

Table 2. Observed frequencies of *give* and the ditransitive in the ICE-GB (expected frequencies in parentheses; from Stefanowitsch & Gries 2003)[1]

	Verb: *give*	**Other verbs**	**Totals**
Construction: ditransitive	461 (9)	574 (1,026)	1,035
Other clause-level constructions	699 (1,151)	136,930 (136,478)	137,629
Totals	1,160	137,504	138,664

The authors give several reasons for choosing p_{FYE}, two of which (cf. Pedersen 1996) I mention here, a third important one will be mentioned in Section 2.3.

i. exact tests do not make distributional assumptions that corpus data usually violate, such as normality and/or homogeneity of variances (cf. Gries & Stefanowitsch 2004: 101);
ii. because of the Zipfian distribution of words in a construction's slot, any AM one might want to use must be able to handle the small frequencies that characterize Zipfian distributions (Stefanowitsch & Gries 2003: 204) and at the same not be anti-conservative.

For Table 2, the p_{FYE} is a very small p-value (<4.94e–324) or a very large $\log_{10}$ of that p-value (>323.3062) so the mutual attraction between *give* and the ditransitive is very strong. This measure is then computed for every verb type in the ditransitive so that the verbs can be ranked according to their attraction to the ditransitive. This entails that the p-values are mainly used "as an indicator of relative importance" (cf. Stefanowitsch & Gries 2003: 239, n. 6), and virtually all collostructional applications have focused only on the 20 to 30 most highly-ranked words and their semantic characteristics (although no particular number is required).

For the second method, *distinctive collexeme analysis* (cf. Gries & Stefanowitsch 2004a), the 2×2 table is set up differently: A corresponds to a construction (e.g., the ditransitive), $\neg A$ corresponds to a functionally similar construction (e.g., the prepositional dative NP V NP $\text{PP}_{for/to}$), B corresponds to a word (e.g., *give*) occurring in syntactically-defined slots of A, and $\neg B$ corresponds to all other words in the slots/the corpus; cf. Table 3.

Table 3. Observed frequencies of *give* and the ditransitive and the prepositional *to*-dative in the ICE-GB (expected frequencies in parentheses; from Gries & Stefanowitsch 2004)

	Verb: *give*	**Other verbs**	**Totals**
Construction: ditransitive	461 (213)	574 (822)	1,035
Construction: prepositional dative	146 (394)	1,773 (1,525)	1,919
Totals	607	2,347	2,954

Again, this results in a very small p_{FYE} (1.835954e-120) or very large negative logged_{10} p-value (119.7361), indicating that *give*'s preference for the ditransitive over the prepositional dative is strong. Again, one would compute this measure for all x verbs attested at least once in either the ditransitive or the prepositional *to*-dative, rank-order the x verbs according to their preference and strength of preference, and then inspect the, say, top t verbs for each construction.

Other extensions of CA are available and have been used. One, *multiple distinctive collexeme analysis*, extends distinctive collexeme analysis to cases with more than two constructions (e.g., the *will*-future vs. the *going-to* future vs. the *shall*-future vs. present tense with future meaning). Another one, *covarying collexeme analysis*, computes measures for co-occurrence preferences *within* one construction (cf. Gries & Stefanowitsch 2004b).[2]

2.3 Perspective 3: CA and its results, interpretation, and motivation

As outlined above, CA returns ranked lists of (distinctive) collexemes, which are analyzed in terms of functional characteristics. For the ditransitive data discussed above with Table 2, the rank-ordering in (1) emerges:

(1) *give*, *tell*, *send*, *offer*, *show*, *cost*, *teach*, *award*, *allow*, *lend*, *deny*, *owe*, *promise*, *earn*, *grant*, *allocate*, *wish*, *accord*, *pay*, *hand*, …

Obviously, the verbs are not distributed randomly across constructions, but reveal semantic characteristics of the constructions they occupy. Here, the verbs in (1) clearly reflect the ditransitive's meaning of transfer (most strongly-attracted verbs involve transfer), but also other (related) senses of this construction (cf. Goldberg's 1995: Ch. 5): (non-)enablement of transfer, communication as transfer, perceiving as receiving, etc.

Similarly clear results are obtained from comparing the ditransitive and the prepositional dative discussed above with Table 3. The following rank-orderings emerge for the ditransitive (cf. (2)) and the prepositional dative (cf. (3)):

(2) *give*, *tell*, *show*, *offer*, *cost*, *teach*, *wish*, *ask*, *promise*, *deny*, …

(3) *bring*, *play*, *take*, *pass*, *make*, *sell*, *do*, *supply*, *read*, *hand*, …

Again, the verbs preferring the ditransitive strongly evoke the notion of transfer, but we also see a nice contrast with the verbs preferring the prepositional dative, which match the proposed constructional meaning of 'continuously caused (accompanied) motion.' Several verbs even provide nice empirical evidence for an iconicity account of the dative alternation as proposed by Thompson & Koide (1987): Verbs such as *bring*, *play*, *take*, and *pass* involve some greater distance between the agent and the recipient (*pass* here mostly refers to passing a ball in soccer), certainly greater than the one prototypically implied by *give* and *tell*.

By now, this method has been used successfully on data from different languages (e.g., English, German, Dutch, Swedish, …) and in different contexts (e.g., constructional description in synchronic data, syntactic 'alternations' (Gilquin 2006), priming phenomena (Szmrecsanyi 2006), second language acquisition (Gries & Wulff 2005, 2009, Deshors 2010), and diachronic language change (Hilpert 2006, 2008). However, while these above examples and many applications show that the CA rankings reveal functional patterns, one may still wonder *why* this works. This question might especially arise given that the most widely-used though not prescribed statistical collostructional measure is in fact a significance test, a *p*-value. Apart from the two mathematical motivations for this *p*-value approach mentioned in the previous section, there is also a more conceptual reason, too.

As all *p*-values, such (logged) *p*-values are determined by both effect and sample size or, in other words, the *p*-value "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 [word] in the [constructional] slot" (Stefanowitsch & Gries 2003: 239, n. 6). For instance, all other things being equal, a percentage of occurrence *o* of a word *w* in *c* (e.g., 40%) is 'upgraded' in importance if it is based on more tokens (e.g., $^{14}/_{35}$) than on less (e.g., $^{8}/_{20}$). This cannot be emphasized enough, given that proponents of CA have been (wrongly) accused of downplaying the role of observed frequencies. CA has in fact been used most often with FYE, which actually tries to afford an important role to observed frequencies: it *integrates* two pieces of important information: (i) how often does something happen — *w*'s frequency of occurrence in *c*, which proponents of observed frequencies rely on — but also (ii) how exclusive is *w*'s occurrence to *c* and *c*'s to *w*. Now why would it be useful to combine these two pieces of information? For instance,

- (i) because "frequency plays an important role for the degree to which constructions are *entrenched* and the likelihood of the production of lexemes in individual constructions (cf. Goldberg 1999)"
  (Stefanowitsch & Gries 2003: 239, n. 6, my emphasis);
- (ii) because we know how important frequency is for learning in general
  (cf., e.g., Ellis 2007);
- (iii) because "collostructional analysis goes beyond raw frequencies of occurrence, […] determining what in psychological research has become known as one of the strongest determinants of prototype formation, namely *cue validity*, in this case, of a particular collexeme for a particular construction"
  (cf. Stefanowitsch & Gries 2003: 237, my emphasis).

In spite of these promising characteristics, Bybee (2010) criticizes CA with respect to each of the three different perspectives outlined above: the goals, the mathematical side, and the results/interpretation of CA. In her claims, Bybee also touches upon the more general point of frequencies vs. AMs as used in many corpus- and psycholinguistic studies. In this paper, I will refute the points of critique by Bybee and discuss a variety of related points of more general importance to cognitive/usage-based linguists.

3. Bybee's points of critique

3.1 Perspective 1: CA and its goals

The most frequent, but by no means only, implementation of CA uses p_{FYE} as an AM, which (i) downgrades the influence of words that are frequent everywhere and (ii) weighs more highly observed relative frequencies of co-occurrence that are based on high absolute frequencies of co-occurrence. Bybee (2010: 97) criticizes this by stating that the "problem with this line of reasoning is that lexemes do not occur in corpora by pure chance" and that "it is entirely possible that the factors that make a lexeme high frequency in a corpus are precisely the factors that make it a central and defining member of the category of lexemes that occurs in a slot in a construction." Using the Spanish adjective *solo* 'alone' as an example, she goes on to say that, for *solo*, "Collostructional Analysis *may* give the wrong results [my emphasis, STG], because a high overall frequency will give the word *solo* a lower degree of attraction to the construction according to this formula" (2010: 98).

3.2 Perspective 2: CA and its mathematics/computation

Bybee (2010: 98) also takes issue with the of the bottom right cell in the 2×2 tables: "Unfortunately, there is some uncertainty about the fourth factor mentioned above — the number of constructions in the corpus. There is no known way to count the number of constructions in a corpus because a given clause may instantiate multiple constructions." Later in the text, however, she mentions that Bybee & Eddington tried different corpus sizes and obtained "similar results" (Bybee 2010: 98).

3.3 Perspective 3: CA and its results, interpretation, and motivation

3.3.1 *The perceived lack of semantics*

Bybee criticizes CA for its lack of consideration of semantics. Specifically, she summarizes Bybee & Eddington (2006), who took "the most frequent adjectives occurring with each of four 'become' verbs as the centres of categories, with semantically related adjectives surrounding these central adjectives depending on their semantic similarity, as discussed above" (Bybee 2010: 98); this refers to Bybee & Eddington's (2006) classification of adjectives occurring with, say, *quedarse*, as semantically related. She then summarizes "[t]hus, our analysis uses both frequency and semantics" whereas "[p]roponents of Collostructional Analysis hope to arrive at a semantic analysis but do not include any semantic factors in their method. Since no semantic considerations go into the analysis, it seems plausible that no semantic analysis can emerge from it" (Bybee 2010: 98).

3.3.2 *The perceived lacks of semantics and discriminatory power*

The above claim is also related to the issue of discriminatory/predictive power. In an attempt to compare Bybee's raw frequency approach to CA, Bybee compares both approaches' discriminability with acceptability judgment data. For two Spanish verbs meaning 'become' (*ponerse* and *quedarse*) and twelve adjectives from three semantic groups (high freq. in *c* with these two verbs, low freq. in *c* but semantically related to the high freq. ones, and low freq. in *c* and semantically unrelated to the high freq. ones), the co-occurrence frequencies of the verbs and the adjectives, the frequency of the adjectives in the corpus, and the collostruction strengths were determined.

As Bybee mentions, frequency and collostruction strength make the same (correct) predictions regarding acceptability judgments for the high-frequency co-occurrences. However, semantically related low-frequency adjectives garner high acceptability judgments whereas semantically unrelated low-frequency adjectives do not. Bybee does not report any statistical analysis, but eyeballing the data seems to confirm this; she states "[o]f course, the Collostructional Analysis cannot make the distinction between semantically related and semantically unrelated since it works only with numbers and not with meaning" (2010: 100). She goes on to say "[t]hus for determining what lexemes are the best fit or the most central to a construction, a simple frequency analysis with semantic similarity produces the best results."

Finally, Bybee criticizes CA in terms of how "many such analyses" handle low-frequency collexemes, which are "ignored" (2010: 101). This is considered a problem because "low-frequency lexemes often show the productive expansion of the category" and "[w]ithout knowing what the range of low frequency, semantically related lexemes is, one cannot define the semantic category of lexemes that can be used in a construction" (p. 101).

3.3.3 *The absence of cognitive mechanisms underlying CA*

From the above claims regarding the relation between frequency, collostruction strength, (semantic similarity), and acceptability judgments, Bybee infers, in agreement with Goldberg's earlier research, that high-frequency lexical items in constructional slots are central to the meaning of a construction. However, she also goes on to claim that

> Gries and colleagues argue for their statistical method but do not propose a cognitive mechanism that corresponds to their analysis. By what cognitive mechanism does a language user devalue a lexeme in a construction if it is of high frequency generally? This is the question Collostructional Analysis must address. (2010: 100f.)

4. Clarifications, repudiations, and responses

This section addresses Bybee's points of critique and other issues. I will show that Bybee's understanding, representation, and discussion of CA does not do the method justice, but the discussion will also bring together a few crucial notions, perspectives, and findings that are relevant to cognitive/usage-based linguists, irrespective of whether they work with CA or not.

4.1 Perspective 1: CA and its goals

There are three main arguments against this part of Bybee's critique. The first is very plain: As cited above, Stefanowitsch & Gries (2003: 217) explicitly state that any AM can be used, one based on a significance test (p_{FYE}, chi-square, *t*, …), one based on some other comparison of observed and expected frequencies (*MI*, MI^2, …), an effect size (Cramer's V/φ, log odds, …), or some other measure (*MinSem*, ΔP, …). For example, Gries (2011, available online since 2006) uses the odds ratio to compare data from differently large corpus parts. Any criticism of CA on these grounds misses its target.

A second, more general counterargument is that the whole point of AMs is to separate the wheat (frequent co-occurrence probably reflecting linguistically relevant functional patterns) from the chaff (co-occurrence at chance level revealing little to nothing functionally interesting). Consider an example on the level of lexical co-occurrence: Whoever insisted on using raw frequencies in contexts alone would have to emphasize that most nouns co-occur with *the* very frequently and that whatever makes *the* occur in corpora is precisely the factor that makes it frequent around nouns. I do not find this particularly illuminating. As a more pertinent example, Bybee's logic would force us to say that the *as*-predicative, exemplified in (4) and discussed by Gries, Hampe & Schönefeld (2005), is most importantly characterized *not* by *regard* (the verb with the highest collostruction strength), but by *see* and *describe*, which occur more often in the *as*-predicative than *regard* (and maybe by *know*, which occurs nearly as often in the *as*-predicative as *regard*). Given the semantics of the *as*-predicative and the constructional promiscuity and semantic flexibility of especially *see* and *know*, this is an unintuitive result; cf. also below.

(4) a. V $\text{NP}_{\text{Direct Object}}$ *as* complement constituent
b. I never saw myself as a costume designer
c. Politicians regard themselves as being closer to actors

It is worth pointing out that the argument against 'testing against the null hypothesis of chance co-occurrence' is somewhat moot anyway. No researcher I know believes words occur in corpora randomly just as no researcher analyzing experimental data believes subjects' responses are random — of course they don't and aren't: if they did, what would be the point of any statistical analysis, with AMs *or* frequencies? With all due recognition of the criticisms of the null hypothesis significance testing paradigm, this framework has been, and will be for the foreseeable future, the predominant way of studying quantitative data — this does not mean the null hypothesis of chance distribution is always a serious contender. Plus, *even if* null hypothesis testing were abandoned, this would still not constitute an argument against AMs because there are AMs not based on null hypothesis frequencies and the most promising of these, ΔP, is in fact extremely strongly correlated with p_{FYE}. Lastly, regardless of which AM is used to downgrade words that are frequent everywhere, *all of them* recognize it is useful to consider not just the raw observed frequency of word *w* in context *c* but also the wider range of *w*'s uses. That is, users of AMs do not argue that the observed frequency of *w* in *c* is unimportant — they argue that it *is* important, as is *w*'s behavior elsewhere. It is surprising that this position could even be criticized from a(n) usage-/exemplar-based perspective, something to which I will return below.

The final counterargument is even more straightforward: Recall that CA involves a normalization of frequencies against corpus size (for CA) or constructional frequencies (for DCA). But sometimes one has to compare 2+ constructions, as in Gries & Wulff (2009), who study *to*/*ing*-complementation (e.g., *he began to smoke* vs. *he began smoking*). They find that *consider* occurs 15 times in both constructions. Does that mean that *consider* is equally important to both? Of course not: the *to*-construction is six times as frequent as the *ing*-construction, which makes it important that *consider* 'managed to squeeze itself' into the far less frequent *ing*-construction as often as into the far more frequent *to*-construction. An account based on frequencies alone could miss that obvious fact — CA or other approaches perspectivizing the observed frequencies of *w* in *c* against those of *w* and/or *c* do not.

4.2 Perspective 2: CA and its mathematics/computation

Let us now turn to some of the more technical arguments regarding CA's input data and choice of measure.

4.2.1 *The issue of the corpus size*

Let us begin with the issue of Bybee's "fourth factor", the corpus size in constructions. Yes, an exact number of constructions for a corpus cannot easily be generated because

i. "a given clause may instantiate multiple constructions" (Bybee 2010: 98);
ii. researchers will disagree on the number of constructions a given clause instantiates;
iii. in a framework that does away with a separation of syntax and lexis, researchers will even disagree on the number of constructions a given word instantiates.

However, this is much less of a problem than it seems. First, this is a problem nearly all AMs have faced and addressed successfully. The obvious remedy is to choose a level of granularity close to the one of the studied phenomenon. For the last 30 years collocational statistics used the number of lexical items in the corpus as *n*, and collostructional studies on argument structure constructions used the number of verbs. Many CA studies, none of which are cited by Bybee or other critics, have shown that this yields meaningful results with much predictive power (cf. also Section 4.2.2 below).

Second, CA rankings are remarkably robust. Bybee herself pointed out that different corpus sizes yield similar results, and a more systematic test supports that. I took Stefanowitsch & Gries's (2003) original results for the ditransitive construction and increased the corpus size from the number used in the paper by a factor of ten (138,664 to 1,386,640), and I decreased the observed frequencies used in the paper by a factor of 0.5 (with *n*'s=1 being set to 0 / omitted). Then I computed four CAs:

- one with the original data;
- one with the original verb frequencies but the larger corpus size;
- one with the halved verb frequencies and the original corpus size;
- one in which both frequencies were changed.

In Figure 1, the pairwise correlations of the collostruction strengths of the verbs are computed (Spearman's rho) and plotted. The question of which verb frequencies and corpus size to use turns out to be fairly immaterial: Even when the corpus size is de-/increased by one order of magnitude and/or the observed frequencies of the words in the constructional slots are halved/doubled, the overall rankings of the words are robustly intercorrelated (all rho>0.87). Thus, this 'issue' is unproblematic when the corpus size is approximated at some appropriate level of granularity and, trivially, consistently, in one analysis.

4.2.2 *The distribution of p_{FYE}*

Another aspect of how CA is computed concerns its 'response' to observed frequencies of word *w* in construction *c* and *w*'s overall frequency. Relying on frequencies embodies the assumption that effects are linear: If something is observed twice as often as something else (in raw numbers or percent), it is, unless another

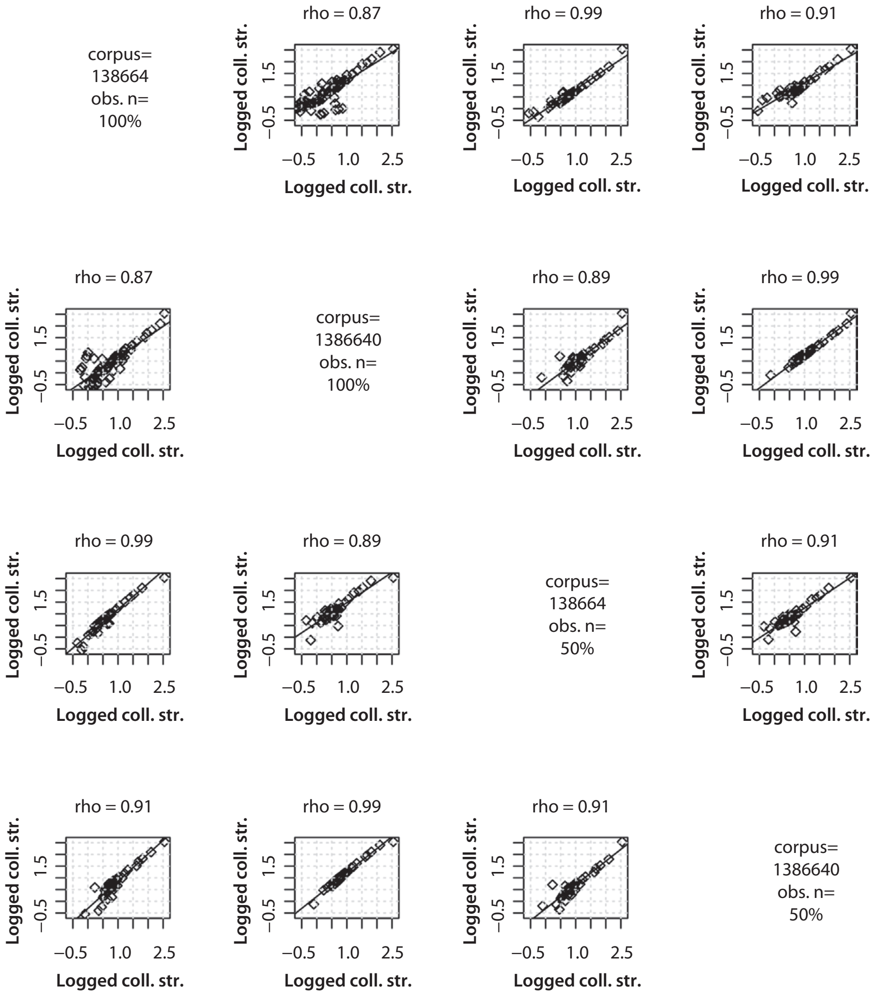


Figure 1. Pairwise comparisons between (logged) collostruction values, juxtaposing corpus sizes (138,664 and 1,386,640) and observed frequencies (actually observed ones and values half that size, with *n*'s=1 being omitted)

transformation is applied, two times as important/entrenched/… However, many effects in learning, memory, and cognition are *not* linear:

- the power law of learning (cf. Anderson 1982, cited by Bybee herself);
- word frequency effects are logarithmic (cf. Tryk 1986);
- forgetting curves are logarithmic (as in priming effects; cf. Gries 2005, Szmrecsanyi 2006), …

Given such and other cases *and* Bybee's emphasis on domain-general processes (which I agree with), it seems odd to rely on frequencies, which have mathematical characteristics that differ from those of many general cognitive processes. It is therefore useful to briefly discuss how frequencies, collostruction strengths, and other measures are related to each other, by exploring systematically-varied artificial data and authentic data from different previous studies.

As for the former, it is easy to show that the AM used in most CAs, p_{FYE}, is not a straightforward linear function of the observed frequencies of words in constructions but rather varies as a function of *w*'s frequency in *c* as well as *w*'s and *c*'s overall frequencies, as Figure 2 partially shows for systematically varied data. The frequency of *w* in *c* is on the *x*-axis, different overall frequencies of *w* are shown in differently grey-shaded points/lines and with numbers, and $-\log_{10} p_{\text{FYE}}$ is shown on the *y*-axis.

I am not claiming that logged p_{FYE}-values are the best way to model cognitive processes — for example, a square root transformation makes the values level off more like a learning curve — but clearly a type of visual curvature we know from many other cognitive processes is obtained. Also, p_{FYE} values are highly correlated with statistics we know *are* relevant in cognitive contexts and that may, therefore, serve as a standard of comparison. Ellis (2007) and Ellis & Ferreira-Junior (2009: 198 and passim) discuss a uni-directional AM called ΔP, which has been used successfully in the associative-learning literature. Interestingly for the data represented in Figure 2, the correlation of p_{FYE} with $\Delta P_{\text{word-to-construction}}$ is

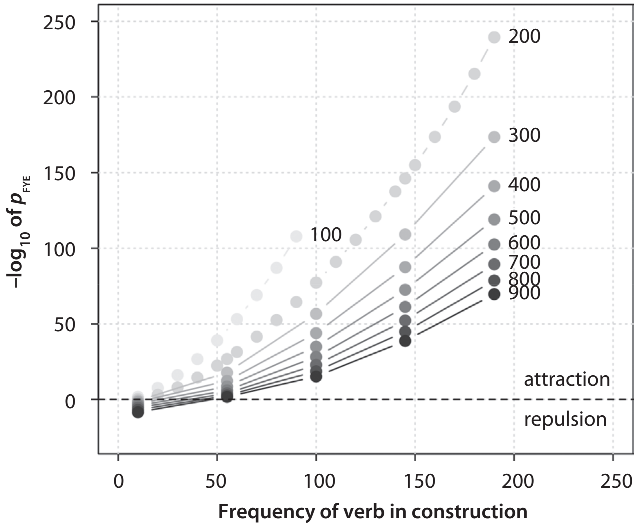


Figure 2: The interaction between the frequency of *w*, the overall frequencies of *w* and *c*, and their collostruction strengths

extremely significant ($p<10^{-15}$) and very high (rho=0.92) whereas the correlations of the observed frequencies or their logs with $\Delta P_{\text{word-to-construction}}$ are significant ($p<10^{-8}$) but much smaller (rho=0.65). Again, p_{FYE} is not necessarily 'the optimal solution', but it exhibits appealing theoretical characteristics ([transformable] curvature, high correlations with measures from learning literature, responsiveness to frequency) that makes one wonder how Bybee can just dismiss them.

Let us now also at least briefly look at authentic data, some here and some further below (in Section 4.3.2). The first result is based on an admittedly small comparison of three different measures of collostruction strengths: For the ditransitive construction, I computed three different CAs, one based on $-\log_{10} p_{\text{FYE}}$, one on an effect size (logged odds ratio), and one on Mutual Information (*MI*). Consider the three panels in Figure 3 for the results, where the logged frequencies of the verbs in the ditransitive are on the *x*-axes, the three AMs are on the *y*-axes, and the verbs are plotted at the *x*/*y*-values reflecting their frequencies and AM values. The correlation between the frequencies and AMs is represented by a polynomial smoother and on the right, I separately list the top 11 collexemes of each measure.

Comparing these results to each other and to Goldberg's (1995) analysis of the ditransitive suggests that, of these measures, p_{FYE} performs best: Starting on the right, *MI*'s results are suboptimal because the prototypical ditransitive verb, *give*, is not ranked highest (let alone by a distinct margin) but only third, and other verbs in the top five are, while compatible with the ditransitive's semantics, rather infrequent and certainly not ones that come to mind first when thinking of the ditransitive. The log odds ratio fares a bit better because *give* is the strongest collexeme, but otherwise the problems are similar to *MI*'s ones.

The p_{FYE}-values arguably fare best: *give* is ranked highest, and by a fittingly huge margin. The next few verbs are intuitively excellent fits for the polysemous ditransitive and match all the senses Goldberg posited: the metaphor of communication as transfer (*tell*), caused reception (*send*), satisfaction conditions implying transfer (*offer*), the metaphor of perceiving as receiving (*show*), etc.; cf. Stefanowitsch & Gries (2003: 228f.) for more discussion. Note also that p_{FYE} also exhibits a behavior that should please those arguing in favor of raw observed frequencies: As the polynomial smoother shows, it is p_{FYE} that is most directly correlated with frequency. At the same time, and this is only a *prima facie* piece of evidence, it is also the p_{FYE}-values whose values result in a curve that has the Zipfian shape that one would expect for such data (given Ellis & Ferreira-Junior's (2009) work (cf. also below).

Finally, there is Wiechmann's (2008) comprehensive study of how well more than 20 AMs predict experimental results regarding lexico-constructional co-occurrence. Raw co-occurrence frequency scores rather well but this was in part because several outliers were removed. Crucially, p_{FYE} ended up in second place and the first-ranked measure, Minimum Sensitivity (*MS*), is theoretically problematic.

Using the notation of Table 1, it is computed as shown in (5), i.e. as the minimum of two conditional probabilities:

(5) $$MS = min\left(\frac{n_{A\&B}}{n_A}, \frac{n_{A\&B}}{n_B}\right) = min(p(word|construction), p(construction|word))$$

One problem here is that some collexemes' positions in the ranking order will be due to *p*(word|construction) while others' will be due to *p*(construction|word). Also, the value for *give* in Table 2 is 0.397, but that does not reveal which conditional probability that value is — *p*(word|construction) or *p*(construction|word). In fact, this can lead to cases where two words get the same *MS*-value, but in one case it is *p*(word|construction) and in the other it is *p*(construction|word). This is clearly undesirable, which is why p_{FYE}, while 'only' second, is more appealing. As an alternative, a uni-directional measure such as Δ*P* is more useful (cf. Gries to appear).

4.3 Perspective 3: CA and its results, interpretation, and motivation

4.3.1 *The perceived lacks of semantics*

I find it hard to make sense of Bybee's first objection to CA, the alleged lack of consideration of semantics discussed in Section 3.3: (i) her claim appears to contradict the exemplar-model perspective that permeates both her whole book and much of my own work; (ii) it does not engage fully with the literature; (iii) it is based on a partial representation of CA, and so it is really arguing against a straw man.

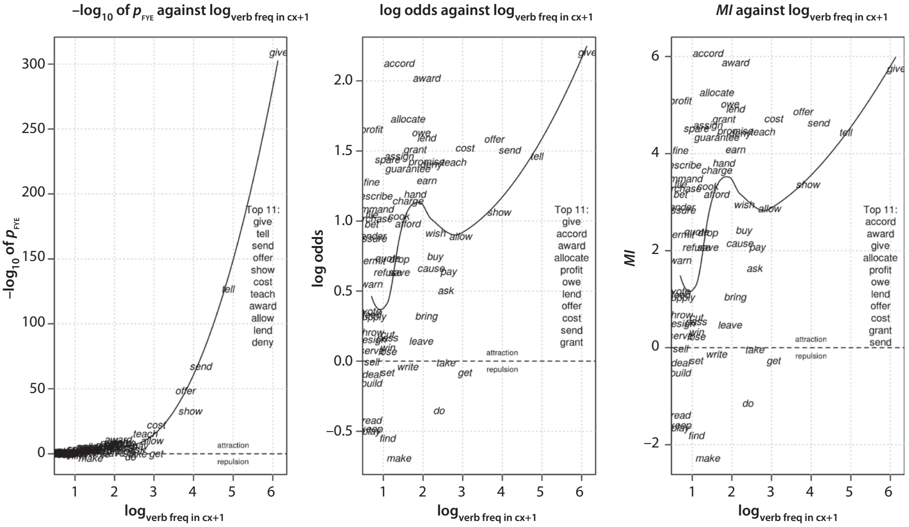


Figure 3. Output scores for the ditransitive of three different AMs (left: p_{FYE}, middle: log odds ratio, right: *MI*)

As for (i), Bybee's statement that "[s]ince no semantic considerations go into the analysis, it seems plausible that no semantic analysis can emerge from it" is false. There is a whole body of work in, e.g., computational (psycho)linguistics where purely frequency-based distributional analyses reveal functionally interpretable clusters. Two classics are Redington, Chater & Finch (1998) and Mintz, Newport & Bever (2002). Both discuss how multidimensional distributional analyses of co-occurrence frequencies reveal clusters that resemble something that, in cognitive linguistics, is considered to have semantic import, namely parts of speech. And even if one did *not* postulate a relation between parts of speech and semantics, both reveal that something can emerge from a statistical analysis (parts of speech) that did not enter into the analysis. Even more paradoxically, it is a strength of exactly the type of usage-/exemplar-based models that Bybee and I both favor that they can explain such processes as the emergence of categories of any kind from processing and representing vast numbers of usage events in multidimensional memory space.

As for (ii), it is even less clear how anyone can imply having read CA studies but claim that collostructional results do not reveal semantic patterns. For example, there are the (discussions of the) lists of collexemes presented in Stefanowitsch & Gries (2003) — recall the ditransitive and the dative alternation from Section 2.3 above — plus there are many other studies aside from Stefanowitsch & Gries (2003) and Gries & Stefanowitsch (2004) — cf. Sections 4.2.2 and 4.3.2 for many examples — nearly all of which have discussed at length functional patterns in the top-ranked collexemes. This lack of engagement with the literature extends even to the CA work speaking most directly to this question: Gries & Stefanowitsch (2010), first presented in 2004 and available online since 2006, clustered the first verbs in the *into*-causative (cf. (6)) based on the *ing*-verbs,[3] and the verbs in the *way*-construction (cf. (7)) based on the prepositions.

(6) a. V NP$_{\text{Direct Object}}$ *into* V-*ing*
b. He tricked her into believing him.
c. They talked you into giving up.

(7) a. V [$_{\text{Direct Object}}$ POSS *way*] PP
b. She fought her way to the stage.
c. He argued his way out of the situation.

Specifically, for each construction they computed a table with all verbs in the construction in the rows, the *ing*-verbs (for the *into*-causative) or the prepositions (for the *way*-construction) in the rows, and the collostructional strengths in the cells. Then, the verbs in the rows (for each construction) were clustered on the basis of the collostructional preferences in the columns using a hierarchical cluster analysis and the resulting tree plot was interpreted in terms of which verbs were grouped together based on similar preferences. These cluster analyses, into which semantics did not enter as data, produced clear semantic patterns. For the *into*-causative, the cluster analysis revealed groups of (more) physical force verbs, of provoking, of trickery, of verbs providing positive stimuli, and of verbs providing negative stimuli. For the *way*-construction, the clustering revealed a cluster of two highly frequent all-purpose verbs, again a group of (more) physical force verbs, and three different clusters reflecting different kinds of slow motion.

In sum, the statement that "[s]ince no semantic considerations go into the analysis, it seems plausible that no semantic analysis can emerge from it" can only be upheld by ignoring both the distributional linguistics literature that Bybee is otherwise sympathetic towards and the specific collostructional literature that she means to criticize and that shows the opposite.

As for (iii), Bybee's comparison of her and Eddington's approach and collostructional data is misleading. Recall the four-step characterization of CA in Section 2.1. On that level of abstraction, Bybee and Eddington's approach consists of the following steps:

- generating a concordance of two words in question (*ponerse* and *quedarse*);
- retrieving frequency data for twelve adjectival collocates of each verb;
- carefully categorizing the adjectives on the basis of their semantic characteristics and frequencies.

Bybee then compares the results of her full-fledged, linguistically informed analysis *not* to the results of an equally full-fledged CA — she compares them to nothing more than the result of applying only step (ii) of a full-fledged CA, as represented in Table 4, which, of course, delivers results that do not have academic merit. To have offered a genuine comparison, Bybee should have computed collostruction strengths of all verbs, not just a small selection and in particular not

Table 4. Bybee's 'Collostructional Analysis'

Step	Real CA as per Section 2.1	Bybee's caricature of a CA
(i)	retrieve all collexeme types	-[4]
(ii)	compute all their collostruction strengths	compute collostruction strengths for 24 adjectives that were the result of her analysis and whose low-frequency items are hapaxes or not attested at all (!)
(iii)	rank-order all collexeme types acc. to their strengths	-
(iv)	analyze the top *n* collexemes semantically / functionally	-

a selection of collexemes occurring maximally once — only then could she have computed the intended rank-ordering which takes high frequencies into consideration and allows for the follow-up semantic analysis of highly-ranked collexemes that many studies have offered. Bybee compares her full analysis to only the numerical output of what Bybee calls a CA rather than the semantic classes of the top-ranked words of a real CA …[5]

4.3.2 *The perceived lacks of semantics and discriminatory power*

There are a number of empirical studies which support CA and undermine Bybee's arguments, which she appears not to have engaged with, especially Gries, Hampe & Schönefeld (2005), although it appears in her list of references. As mentioned above, Gries, Hampe & Schönefeld (2005) studied the *as*-predicative by means of a CA. They then ran a factorial sentence-completion experiment in which subjects were presented with sentence fragments ending in one of a set of verbs. These verbs were from eight groups that resulted from all combinations of three independent binary variables: COLLSTR (high vs. low), FREQCX (high vs. low), and VOICE (the voice of the sentence fragment: active vs. passive); a second re-analysis of the data also included FAITH (p(construction|verb)) as a covariate. ANOVAs of both analyses revealed highly significant effects of COLLSTR (also with the highest effect size) and insignificant and very weak effects of FREQCX. A follow-up study, Gries, Hampe & Schönefeld (2010, first presented 2004 and available online since 2006) revisited the *as*-predicative with a self-paced reading time study. Subjects' reading times on words after *as* were measured to determine whether the (dis)preference of a verb for the *as*-predicative would speed up/slow down reading processes when an *as*-predicative is encountered or not. While the result for COLLSTR very narrowly missed standard levels of significance (p=0.0672, effect size=0.014), this result would have been significant in a justifiable one-tailed test,[6] and FREQCX yielded insignificant/weak results (p=0.293, effect size=0.005).

Bybee also ignores other studies that, while not primarily devoted to similar comparisons, still speak to the issue:

- Gries & Wulff (2005, 2009) find strong correlations between collostruction strengths and experimentally-obtained sentence completions from advanced L2 learners of English;
- Ellis & Ferreira-Junior (2009) find that frequency of learner uptake is predicted by frequency of occurrence, but more so by p_{FYE} and ΔP;
- both Gries (2005) and Szmrecsanyi (2006) find strong correlations between verbs' collostruction strengths and priming effects observed in different corpora and for different constructions.

In sum, Bybee systematically chooses to not mention results of even a single study with experimental and/or corpus-based data running counter to her claims, but even a cursory glance at the literature shows that the picture is the opposite of the one she painted or, at least, much more complicated.

Bybee's final point of critique regarding low-frequency collexemes is only too easy to counter. No one ever said low-frequency collexemes *should* be ignored or cannot be revealing. A CA is based on the very fact that *all* collexemes are included — the fact that most studies have focused on the top collexemes that are functionally most revealing does not mean weakly-attracted or repelled collexemes should not be studied, and the software that most CAs have used offers estimates of collostruction strengths for unattested words.

4.3.3 *The absence of cognitive mechanisms underlying CA*

Similarly straightforward to refute is the implication that CA does not come with a cognitive account of the data. First, given the strong (experimental and otherwise) support of collostruction strength in many studies that all adopt a cognitive-linguistic/usage-based framework, it is surprising there should be a special need for a cognitive underpinning in addition to what all these studies are based on anyway.

Second, the earliest studies make it very clear what their cognitive underpinning is. In Section 2.3 above, I already provided several quotes (from the studies Bybee refers to) to illustrate the CA position: Ultimately, collostruction strengths are based on (i) the conditional probabilities $p(\text{word}|\text{construction})$ and $p(\text{construction}|\text{word})$, which are related to notions of cue validity, cue reliability (cf. Goldberg 2006: Ch. 5–6 and Stefanowitsch to appear), associative learning measures such as ΔP, and prototype formation, and (ii) the frequencies that give rise to the probabilities, which are correlated with entrenchment. Put yet another way: "it is assumed […] that the statistical associations found in the data are reflected in psychological associations in the mind of the language user" (Stefanowitsch 2006: 258).

5. Towards a new empirical perspective and its theoretical implications

5.1 A cline of co-occurrence complexity and its motivations/implications

So far this paper has been concerned with documenting how CA is, contrary to Bybee's claims, a good tool for the analysis of co-occurrence data from corpora. However, it is now worth returning in more detail to two questions that were discussed only briefly above: (i) why exactly does CA provide the (relatively) good results that it does and (ii) what is the cognitive mechanism that it reflects/assumes? In what follows, I will discuss these issues in detail because a more elaborate treatment of them has profound implications on how (different kinds of) data inform cognitive-linguistic theory and establish connections to other theoretical approaches. To explore the answers to these questions and their implications, I will outline a cline of co-occurrence complexity of how to study corpus data and, as this cline is built up, discuss how each step of increased methodological complexity is motivated theoretically; ultimately, this build-up will result in what I think is a necessary clarification of what a usage-/exemplar-based approach entails both in terms of data and theoretical notions such as *construction*.

5.1.1 *Approach 1: Raw frequencies/percentages*

As a first step on the co-occurrence cline, let's look at a raw frequency/percentage type of approach, which is represented in Figure 4: "*w*1", "*w*2", etc. and "*c*1" stand for 'word 1', 'word 2', etc. (e.g., *give*, *tell*, etc.) and 'construction 1' (e.g., the ditransitive) respectively.

This information is often easy to obtain and can be useful in a variety of applications as Bybee and others have shown. As argued above, this approach is also extremely restrictive in that it adopts a very limited view of the more complex reality of use. Among other things, it focuses on only one context, *c*1, and does not take into consideration uses of *w*1, *w*2, etc. outside of *c*1 into consideration, something which the next approach, AMs, does.

5.1.2 *Approach 2: Association measures*

As argued in detail above, AMs consider uses of *w*1, *w*2, … outside of *c*1, cf. Figure 5. The bold Figures **80**, **60**, and **40** here correspond to those in Figure 4; the italics will be explained below.

	c1
*w*1	**80**
*w*2	**60**
*w*3	**40**
…	…

Figure 4. Approach 1: Observed frequencies of words 1-*x* in construction 1

	*c*1	other	Sum
*w*1	**80**	*200*	**280**
other	1000	…	…
Sum	**1080**	…	sum

	*c*1	other	Sum
w1	**60**	*310*	**370**
other	1020	…	…
Sum	**1080**	…	sum

	*c*1	other	Sum
w1	**40**	*420*	**460**
other	1040	…	…
Sum	**1080**	…	sum

Figure 5. Approach 2: AMs for occurrences of words 1–3 (of *x*) in construction 1

Obviously, Figure 5 illustrates a more comprehensive approach than Figure 4: This is true in the trivial sense that all the information in Figure 4 is also present in Figure 5, plus more, namely the token frequencies of the words *w*1–3 outside of *c*1 and the frequency of *c*1. But this is also true in the sense that this is the CA approach that, as discussed above, proved superior in terms of explaining completion preferences, reading times, and learner uptake.

It is probably fair to say that, in general, approach 2 is one of the more sophisticated ways in which co-occurrence data are explored in contemporary usage-based linguistics. However, while I have been defending just this AM approach against the even simpler approach of Figure 4, it is still only a caricature of what is necessary, as we will see in the next section.

5.1.3 *Approach 3: Full cross-tabulation*

Figure 6 shows the next step on the cline, a full cross-tabulation of words and their uses in contexts/constructions.

Again, this approach is more comprehensive than the preceding ones; it contains all their information, and more. This additional information is very relevant within usage-based theory and should, therefore, also figure prominently in usage-based analyses of data.

First, approach 3 provides crucial information on type frequencies that both previous approaches miss. Approach 1 only stated that *w*1 occurs in *c*1; approach 2 stated that *w*1 occurs in *c*1 but also elsewhere and that *c*1 occurs with *w*1 and also elsewhere. Approach 3, however, zooms in on the 200 elsewhere-uses of *w*1 and the 1000 elsewhere-uses of *c*1 (italicized in Figure 5) by revealing, for instance,

	*c*1	*c*2	*c*3	*c*4	*c*5	*c*6	*c*7-15	Sum	types	*H*
*w*1	**80**	*90*	*45*	*35*	*25*	*5*	*0*	**280**	6	2.26
*w*2	**60**	*0*	*310*	*0*	*0*	*0*	*0*	**370**	2	0.639
*w*3	**40**	*30*	*30*	*30*	*30*	*30*	*270*	**460**	15	3.902
*w*4	40	407	1	1	1	1	9	460	15	0.713
*w*5	40	420	0	0	0	0	0	460	2	0.426
*w*6	40	1	407	1	1	1	9	460	15	0.713
*w*7	40	0	420	0	0	0	0	460	2	0.426
*w*8-20	…	…	…	…	…	…	…	…	…	…
Sum	**1080**	948	1213	…	…	…	…	sum	15	…
types	…	…	…	…	…	…	…	…	…	…
H	…	…	…	…	…	…	…	…	…	…

Figure 6. Approach 3: Cross-tabulation of words *w*1–20 and constructions *c*1–15. The row/column 'types' represents the number of constructions/words a word/construction is attested with. The row/column *H* represents the uncertainty/entropy of the token distributions.[7,8]

that *w*1 occurs in 6 out of the 15 constructions; analogously for the 310 and 420 elsewhere-uses of *w*2 and *w*3 in 2 and all 15 constructions respectively, etc.

This kind of type-frequency information is already important for many pertinent reasons. On the one hand, there are results showing that type frequencies are relevant to acquisition, and recent studies on a new AM that incorporates type frequencies — gravity — have yielded very promising results (cf. Daudaravičius & Marcinkevičienė 2004, Gries 2010a). However, there is an even more important theoretical motivation, namely how type frequencies tie in with psycholinguistic/cognitive-psychological theories. Consider, for instance, the so-called fan effect, which is "[s]imply put, the more things that are learned about a concept [the more factual associations fan out from the concept], the longer it takes to retrieve any one of those facts" (Radvansky 1999: 198).[9] While the analogy is admittedly crude, the first clause can be seen as involving the number of connections (i.e., a kind of type frequency) between, say, a construction and the range of words that can be used in it (or a word and the range of constructions it can be used in). Following this analogy, in a cognitive architecture such as Anderson's ACT-R theory, the strength of activation S_{ji} between a source of activation *j* and a fact *i* is dependent on the log of the fan: "activation […] will decrease as a logarithmic function of the fan associated with the concept. […] the strengths of associations decrease with fan because the probability of any fact, given the concept, decreases with fan" (Anderson & Reder 1999: 188). For the association of a word to constructions, this would mean that the strength of the word's associations will be affected by the number of constructions to which it is connected, and vice versa for the association of a construction to words, which shows that the number of types with which words/constructions occur is, contra approach 1, undoubtedly cognitively relevant. In fact, as I will discuss now, it is not just this type frequency that is important.

Second, approach 3 provides not just the type frequencies just discussed, but also the type-token distributions: Not only do we now know that *w*1 appears in *c*1 and in 5 other constructions — we also know with which (italicized) frequencies (80 in *c*1, plus 90, 45, 35, 25, and 5 instances in *c*2–6); analogously for the other words *and* the other constructions. This raises an important issue which most usage-based theorizing discusses very little: Is there any reason to regard this level of resolution as relevant especially given Bybee's (2010: 100f.) question, "[b]y what cognitive mechanism does a language user devalue a lexeme in a construction if it is of high frequency generally?" In approach 1, of course, the question of 'devaluing' does not arise because one does not have to consider where, other than in construction *c*1, word *w*1 occurs. However, by insisting that the distribution of a word *w*1 outside of the construction *c*1 is irrelevant (cf. p. 100) and that only the frequency of *w* in *c* is needed, Bybee and other proponents of approach 1 run into a huge problem. Not only have we seen above that type frequencies are already relevant to a truly cognitive approach, but Bybee (2010: 89) herself also approvingly states "Goldberg 2006 goes on to argue that in category learning in general a centred, or low variance, category is easier to learn." This correctly emphasizes the importance of type-token distributions — but her own approach 1 does not incorporate the very type frequencies and type-token distributions which allow usage-based theorists to talk about 'centred, or low variance, categories' in the first place.

As another example of the importance of type-token distributions, consider Goldberg, Casenhiser, & Sethuraman's (2004) learning experiment: Subjects heard the same number of novel verbs (type frequency: 5), but with two different distributions of 16 tokens. These different token distributions — a balanced condition of 4-4-4-2-2 (with an entropy of H=2.25) and a skewed lower-variance condition of 8-2-2-2-2 (H=2). The more skewed distribution was learned significantly better, but proponents of a radical approach 1 cannot explain this very well since both conditions involved 16 tokens. Proponents of approach 3, on the other hand, can explain this result perfectly with reference to the lower entropy/uncertainty of the skewed distribution; in a similar vein, it is such type-token distributions that help explain the issue of preemption.

Similar examples of how such more comprehensive co-occurrence information is useful abound. The classics of Redington, Chater & Finch (1998) and Mintz, Newport & Bever (2002) are based on similar co-occurrence matrices (based on bigram frequencies, however), as is Latent Semantic Analysis. McDonald & Shillcock (2001: 295) demonstrate that:

> Contextual Distinctiveness (CD), a corpus-derived word recognition summary measure of the frequency distribution of the contexts in which a word occurs [based on H_{rel}, STG] […] is a significantly better predictor of lexical decision latencies than occurrence frequency, suggesting that CD is the more psychologically relevant variable.

Recchia, Johns & Jones (2008: 271f.) summarize their study:

> The results […] suggest that lexical processing is optimized for precisely those words that are most likely to be required in any given situation. […] context variability is potentially a more important variable than is frequency in word recognition and memory access.

Raymond & Brown (2012) find that word frequency plays no role for reduction processes once contextual co-occurrence factors are taken into consideration; Baayen (2010) discusses comprehensive evidence for the relevance of rich contextual and entropy-based measures. Thus, in addition to the many problems of Bybee's argumentation addressed above, there is a large number of theoretical approaches and empirical studies in corpus and psycholinguistics that powerfully converge in their support of a usage-based approach that invokes much more contextual information than the CA-type of approach 2, let alone approach 1 — at the very least, we need type frequencies of co-occurrence of words and constructions and their type-token distributions.

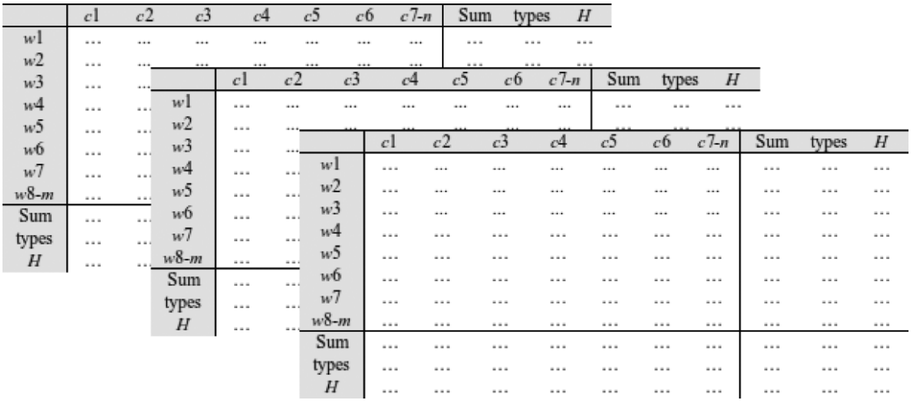
5.1.4 *Approach 4: Dispersion of (co-)occurrence*

In some sense, unfortunately, the two-dimensional cross-tabulation of Figure 6 is still not sufficient: What is missing is how widespread in language use a particular (co-)occurrence is, a notion that is known as dispersion in corpus linguistics (cf. Gries 2008). Essentially we need a three-dimensional approach in which cross-tabulations such as Figure 6 are obtained for a third dimension, namely one containing 'corpus parts,' which could correspond to registers/genres or any other potentially relevant distinction of usage events; cf. Figure 7.

Dispersion is relevant because frequent co-occurrence or high attractions are more important when they are attested in many different registers or situations or other types of usage events, which affects how associations between linguistic elements are discovered/learned:

> Given a certain number of exposures to a stimulus, or a certain amount of training, learning is always better when exposures or training trials are distributed over several sessions than when they are massed into one session. This finding is extremely robust in many domains of human cognition. (Ambridge et al., 2006: 175)

Stefanowitsch & Gries (2003) find that the verbs *fold* and *process* are both relatively frequent in the imperative, occurring 16 and 15 out of 32 and 44 times, respectively, in the imperative, and are highly attracted to it (with collostruction values of 21 and 16.7, respectively). However, both verbs occurred in the imperative



	*c*1	*c*2	*c*3	*c*4	*c*5	*c*6	*c*7-*n*	Sum	types	*H*
*w*1	...	...	...	...	...	...	...	...	...	...
*w*2	...	...	...	...	...	...	...	...	...	...
*w*3	...	...	...	...	...	...	...	...	...	...
*w*4	...	...	...	...	...	...	...	...	...	...
*w*5	...	...	...	...	...	...	...	...	...	...
*w*6	...	...	...	...	...	...	...	...	...	...
*w*7	...	...	...	...	...	...	...	...	...	...
*w*8-*m*	...	...	...	...	...	...	...	...	...	...
Sum	...	...	...	...	...	...	...	...	...	...
types	...	...	...	...	...	...	...	...	...	...
H	...	...	...	...	...	...	...	...	...	...

Figure 7. Approach 4: Cross-tabulation of words *w*1-*m* and constructions *c*1-*n* in (here, 3) different slices/parts of a corpus

in only one of the 500 files of the corpus studied; their dispersion values *DP* (cf. Gries 2008) are >0.99, which indicates their absolutely unrepresentative clumpiness in the corpus, which in turn means their relevance to the imperative should be downgraded especially when compared to *hang on*, which is just as frequent in the imperative but occurs in many more corpus files. Thus, while frequency of (co-)occurrence is related to dispersion — *on the whole*, frequent items will be more dispersed, less frequent items will be more clumpy — this correlation is by no means absolute, and Gries (2010b) shows that dispersion is sometimes a better predictor of reaction times than frequency. Therefore, a cognitively realistic approach should include dispersion and even different word senses.

5.2 Why CA works at all and a brief excursus on Zipf

It is useful to now consider the question of how it is even possible that CA works as well as it does although its inclusion of context, while better than approach 1, is still so impoverished. After all, all it includes is two token frequencies (e.g., 200 and 1000 for *w*1) rather than two type frequencies and their type-token distributions let alone dispersion.

As I see it, CA works as well as it does — and especially so when used with p_{FYE} — for several reasons, most of which are typically not recognized. First,

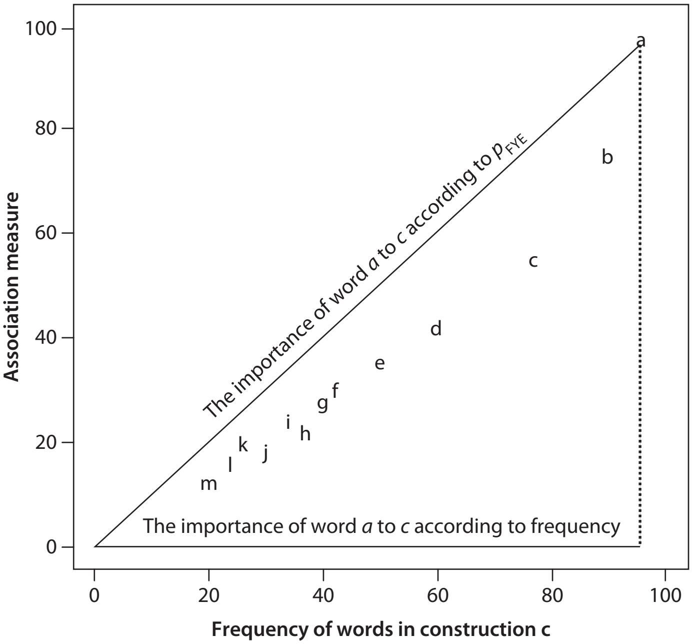


Figure 8. The comparison of a frequency- vs an AM-based approach

because, as Ellis & Ferreira-Junior (2009) show, the correlation of p_{FYE} with the above-mentioned ΔP measure associative learning is high. This is so because CA approximates the type-token distributions of approaches 3 and 4 by including the corresponding token frequencies $n_{w1 \text{ outside of } c1}$ (200) and $n_{c1 \text{ without } w1}$ (1000) — rather than ignoring them as frequencies do — and because, as a *p*-value, p_{FYE} weighs observed percentages of co-occurrence more strongly as the overall *n* of a 2×2 table increases (recall Section 2.3). This logic can be visualized as in Figure 8, which represents the frequencies of words *a* to *m* in construction *c* as well as their attraction to *c*. According to approach 1, the value that reflects how important *a* is for the analysis of *c* is the horizontal line at the bottom, the line from the origin to the *x*-value (the frequency) of *a* in *c*. However, AMs add information (on the *y*-axis) so the value that reflects how important *a* is for the analysis of *c* becomes the line from the origin to *a* in the top right corner; this additional information is one reason why approach 2/CA does often better than approach 1.

There is a second and theoretically more important reason why CA works, and this is concerned with a characteristic of language that is, with some exceptions, topicalized too little in the usage-based approach: the Zipfian distribution that linguistic elements within, say, syntactically-defined slots, exhibit. What makes CA work most of the time is that the 200 and 1000 elsewhere-uses in the above example of *w*1 will be Zipfian-distributed, which in turn means that, especially with high frequencies of a word in a construction, all other uses will be much rarer and thus not distort the data much.

Finally, there is an implication of Zipfian distributions that is little commented on in cognitive/usage-based linguistics: We know that the frequencies of words in constructional slots are Zipfian distributed (Ellis & Ferreira-Junior 2009), we know that skewed low-variance distributions lead to better learning than balanced ones (Goldberg et al. 2004; Goldberg 2006), and we know that type frequency is related to productivity (Goldberg 2006: 99). Goldberg (2006: 89) speculates that this "may involve a type of cognitive anchoring," but I think another (yet not incompatible) perspective is to realize that Zipfian distributions involve less uncertainty than random, uniform, or less Zipfian distributions: The more tokens are accounted for by fewer types, the lower the entropy of the distribution, as is exemplified informally in Figure 9.

Thus, the notion of entropy not only highlights the need to go beyond approaches 1 and 2, but also unites many findings in cognitive-linguistic theorizing under one umbrella. In fact, it unites cognitive-linguistic theorizing with recent approaches in psycholinguistics that study constructional choices on the basis of notions such as surprisal (an information-theoretic operationalization of 'surprise', cf. Jaeger & Snider 2008) and unified information density (Frank & Jaeger 2008) or

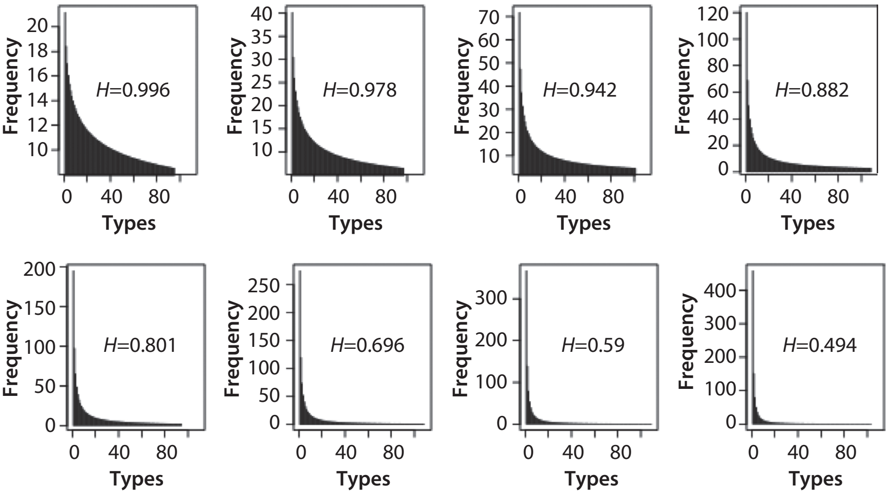


Figure 9. Pseudo-random Zipfian distributions and their entropies

study how category learning and productivity arise from Hebbian learning from Zipfian input with low entropy (cf. Zeldes 2011).

5.3 Towards a refined usage-/exemplar-based definition of construction

In Goldberg (1995: 4), constructions were defined as a form-meaning pair with at least one unpredictable property. In Goldberg (2006: 5), a different definition is proposed: Something unpredictable is no longer a necessary condition — something can also be a construction by virtue of "sufficient frequency." In usage-/exemplar-based models, linguistic/constructional knowledge is conceived of as a high-dimensional space with formal (phonetic, phonological, morphological, syntactic, …) and functional (semantic, pragmatic, discoursal, contextual, …) dimensions. In such a space, exemplars are stored in positions representing their values on these dimensions, and clouds of exemplars with high densities (compared to the space surrounding them) are what corresponds to categories. If a speaker encounters a linguistic token with a particular function in a particular context, then this token is categorized according to its position in the high-dimensional space and will be categorized as a member of the category (point cloud) to which it is most similar (closest).

The above discussion of the cline of co-occurrence complexity and entropy gives rise to a different kind of definition of *construction*. I view a construction as an entropy-reducing spike of a distribution in an area in multidimensional space where formal and functional dimensions intersect. That is, when a point cloud is particularly dense compared to its environment, that means that particular combinations of features (the densest center of the cloud) are more frequent than many others, giving rise to the peak in a Zipfian distribution. An example will help to clarify this rather abstract notion. Consider a child's growing linguistic knowledge as a multidimensional space of formal and functional characteristics before that child has begun to acquire a ditransitive construction. As part of his input, the child hears words such as *give*, *tell*, *show*, *hand*, etc. as verbs in the ditransitive construction, but also in other formal contexts (*give up*, *the show on TV*, *my hand hurts*, …) with different meanings (i.e., functional characteristics). According to the above proposal, the child may begin to form a ditransitive construction when he 'realizes' that *give* does not occur randomly frequently (i.e., with high uncertainty/entropy) in different formal contexts (a.k.a. constructions) and with different meanings but that:

- the distribution of formal contexts with which *give* occurs features a high frequency of ditransitive constructions (plus maybe some other constructions), resulting in a low-uncertainty Zipfian distribution along this dimension;
- the distribution of meanings with which *give* occurs features a high-frequency of 'transfer' meanings (plus maybe some other meanings), giving rise to a low-uncertainty Zipfian distribution along this dimension.

When such an informative confluence of formal and at least one functional characteristic is noticed, an (at first) item-specific construction can emerge, which is then extended more productively as the low-frequency range of uses of the same construction is noticed — the child begins to be able to handle the higher entropy/uncertainty of this distribution — and, for example via Hebbian learning, extends the category.

This perspective helps operationalize Goldberg's notion of "sufficient frequency" more precisely/meaningfully: a frequency is "sufficient" if the frequency of a confluence of one or more formal and one or more functional characteristics has become skewed/Zipfian enough to reduce the uncertainty along the dimensions characterizing the distribution. Note that this means that, for a productive category to emerge, a certain type frequency will be necessary because it is against the background of that type frequency (i.e., many low-frequency bars in any panel of Figure 9) that an entropy-reducing spike (a high-frequency bar in any panel of Figure 9) can be registered. Note also that such a 'realization' of an uncertainty-reducing confluence can of course be facilitated by the salience of the particular confluence in some context, which helps account for instances of fast-mapping and long retention, e.g. when two casual mentions of the nonce-color term *chromium* was sufficient for three- and four-year-old to infer and retain the word's meaning. Crucially, the casual mentions were in a contrastive context ("not blue"), which in the current account simply means that the discoursal context reduced the uncertainty of what *chromium* refers to in semantic space to a degree that children could make the relevant inference.

5.4 Conclusion

While Section 5 has covered a lot of ground, this should not detract from, but reinforce, the realization that the cline of co-occurrence complexity, entropy, and spikes in multidimensional space all point to the same conclusion with regard to corpus data in cognitive/usage-based linguistics — that raw one-dimensional frequencies/percentages are too crude a tool to go the long way we still have to go towards understanding the cognitive and statistical properties of language acquisition, processing, use, and change. No one has summarized it better than Ellis & Ferreira-Junior (2009: 194):

> Raw frequency of occurrence is less important than the contingency between cue and interpretation. Distinctiveness [in multidimensional space, STG] or reliability of form-function mapping is a driving force of all associative learning, […] Contingency, and its associated aspects of predictive value, information gain, and statistical association, have been at the core of learning theory ever since.

Once we add to this perspective truly multidimensional approaches and new developments in distributional learning that can be applied to such informationally rich contexts (cf. Baayen's 2011 paper on a naïve discriminative learning approach inspired by very the same approach by Rescorla and Wagner that Ellis and Ferreira-Junior's work discusses), then we stand a chance of developing better theories for our data — dumbing down our methods and/or ignoring various kinds of converging evidence, on the other hand, will not help.

Notes

* This paper is a revised and extended version of a plenary talk I gave at the 6th International Conference on Construction Grammar in Prague. I thank the audience there, workshop participants and panel discussants at the Freiburg Institute of Advanced Studies, students of my doctoral seminar on psycholinguistics at UCSB, the audience of a Linguistics Colloquium talk at UC Berkeley, and (in alphabetical order) William Croft, Sandra C. Deshors, Adele E. Goldberg, Anatol Stefanowitsch, and Stefanie Wulff for feedback, input, and/or discussion. I also thank two anonymous reviewers and the editors of this special issue for their comments. The usual disclaimers apply.

[1] The expected frequencies are computed as in every contingency table or in chi-square tests for independence. The expected frequency in each cell is the result of row total times column total divided by the sum of all frequencies in the table. For instance, 1035·1160/138,664 ≈ 8.66 ≈ 9.

[2] All of these CA methods (with different AMs) can be computed easily with an interactive R script available at <http://tinyurl.com/collostructions>.

[3] Bybee (2010: 81) quotes Gries et al. (2005) for "verbs occurring in the *int*o-causative," but these are not discussed in that paper (but in Gries & Stefanowitsch 2004b, 2010).

[4] More precisely, it is unclear whether this step was undertaken or not, but no data/analysis is offered of collexemes other than the 24 mentioned and it is possible to compute Bybee's collostruction strengths on the basis of the lexical frequencies of *ponerse*, *quedarse*, and the 24 adjectives.

[5] In fact, all those things are still not the only ones in which her comparison is problematic. For example, she only 'tests' how well the 'high acceptability' judgments are predicted — what about the 'low acceptability' judgments? The real analysis would have also included the 'low acceptability' judgments, which *could* already change the results, because it is well known that low frequency of occurrence does not necessarily mean 'low acceptability'; cf. Stefanowitsch (2005, 2007, 2008) on the relation of collostruction strength to negative acceptability and negative evidence.

[6] A one-tailed test would have been justifiable because the expectation was that high collostruction strengths to the *as*-predicative would not just result in different reading times, but faster ones.

[7] *H*, entropy, is a measure of uncertainty, or dispersion, for categorical data which quantifies how evenly distributed elements are across categories. It ranges from 0 (for perfectly skewed/predictable distributions such as {0, 0, 0, 100}) to $\log_2 n$ (for perfectly equal/unpredictable distributions such as {25, 25, 25, 25}; cf. Gries (2009: 112f.).

[8] Data of this type are of course extremely hard to obtain (especially with a reasonable degree of precision) but see Roland, Dick & Elman (2007) for one recent attempt.

[9] I thank a reviewer for pointing out this connection.

References

Ambridge, Ben, Anna L. Theakston, Elena V. M. Lieven & Michael Tomasello. 2006. The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development* 21(2). 174–193.

Anderson, John R. 1982. Acquisition of cognitive skill. *Psychological Review* 89(4). 369–406.

Baayen, R. Harald. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon* 5(3). 436–461.

Baayen, R. Harald. 2011. Corpus linguistics and naïve discriminative learning. *Brazilian Journal of Applied Linguistics* 11(2). 295–328.

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

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

Daudaravičius, Vidas & Rūta Marcinkevičienė. 2004. Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics* 9(2). 321–348.

Deshors, Sandra C. 2010. *A multifactorial study of the uses of 'may' and 'can' in French-English interlanguage*. University of Sussex dissertation.

Ellis, Nick C. 2007. Language acquisition as rational contingency learning. *Applied Linguistics* 27(1). 1–24.

Ellis, Nick C. & Fernando Ferreira-Junior. 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7. 187–220.

Frank, Austin F. & T. Florian Jaeger. 2008. Speaking rationally: Uniform Information Density as an optimal strategy for language production. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, 939–944.

Gilquin, Gaëtanelle. 2006. The verb slot in causative constructions. Finding the best fit. *Constructions* 1. 1–3.

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

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Goldberg, Adele E., Devin M. Casenhiser & Nitya Sethuraman. 2004. Learning argument structure generalizations. *Cognitive Linguistics* 15(3). 289–316.

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437.

Gries, Stefan Th. 2009. *Statistics for linguistics with R*. Berlin & New York: Mouton de Gruyter.

Gries, Stefan Th. 2010a. Bigrams in registers, domains, and varieties: A bigram gravity approach to the homogeneity of corpora. In *Proceedings of Corpus Linguistics 2009*, University of Liverpool.

Gries, Stefan Th. 2010b. Dispersions and adjusted frequencies in corpora: Further explorations. In Stefan Th. Gries, Stefanie Wulff & Mark Davies (eds.), *Corpus linguistic applications: Current studies, new directions*, 197–212. Amsterdam: Rodopi.

Gries, Stefan Th. 2011. Th. 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: John Benjamins.

Gries, Stefan Th. to appear. 50-something years of work on collocations: What is or should be next … *International Journal of Corpus Linguistics*.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004a. Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9(1). 97–129.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004b. Co-varying collexemes in the *into*-causative. In Michel Achard & Suzanne Kemmer (eds.), *Language, culture, and mind*, 225–236. Stanford, CA: CSLI.

Gries, Stefan Th. & Anatol Stefanowitsch. 2010. Cluster analysis and the identification of collexeme classes. In John Newman & Sally Rice (eds.), *Empirical and experimental methods in cognitive/functional research*, 73–90. Stanford, CA: CSLI.

Gries, Stefan Th., Beate Hampe & Doris Schönefeld. 2005. Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16(4). 635–676.

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

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

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

Harris, Zellig S. 1970. *Papers in structural and transformational linguistics*. Dordrecht: Reidel.

Hilpert, Martin. 2006. Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory* 2(2). 243–257.

Hilpert, Martin. 2008. *Germanic future constructions: A usage-based approach to language change*. Amsterdam: John Benjamins.

Jaeger, T. Florian & Neal Snider. 2008. Implicit learning and syntactic persistence: Surprisal and cumulativity. In Brad C. Love, Ken McRae & Vladimir M. Sloutsky (eds.), *Proceedings of the Cognitive Science Society Conference*, 1061–1066. Washington, DC.

McDonald, Scott A. & Richard C. Shillcock. 2001. Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech* 44(3). 295–323.

Mintz, Toben H., Elissa L. Newport & Thomas G. Bever. 2002. The distributional structure of grammatical categories in speech to young children. *Cognitive Science* 26(4). 393–424.

Pecina, Pavel. 2009. Lexical association measures and collocation extraction. *Language Resources and Evaluation* 44(1–2). 137–158.

Pedersen, Ted. 1996. Fishing for exactness. *Proceedings of the South-Central SAS Users Group Conference (SCSUG-96).* Austin, TX, Oct 27–29.

Radvansky, Gabriel A. 1999. The Fan Effect: A tale of two theories. *Journal of Experimental Psychology: General* 128(2). 198–206.

Raymond, William D. & Esther L. Brown. 2012. Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish. In Stefan Th. Gries & Dagmar S. Divjak (eds.), *Frequency effects in language learning and processing*, 35–52. Berlin & New York: De Gruyter Mouton.

Recchia, Gabriel, Brendan T. Johns & Michael N. Jones. 2008. Context repetition benefits are dependent on context redundancy. *Proceedings of the Annual Conference of the Cognitive Science Society* 30. 267–272.

Redington, Martin, Nick Chater & Steven Finch. 1998. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science* 22(4). 435–469.

Roland, Douglas, Frederick Dick & Jeffrey L. Elman. 2007. Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language* 57(3). 348–379.

Stefanowitsch, Anatol. 2005. New York, Dayton (Ohio), and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory* 1(2). 295–301.

Stefanowitsch, Anatol. 2006. Distinctive collexeme analysis and diachrony: A comment. *Corpus Linguistics and Linguistic Theory* 2(2). 257–262.

Stefanowitsch, Anatol. 2007. Linguistics beyond grammaticality. *Corpus Linguistics and Linguistic Theory* 3(1). 57–71.

Stefanowitsch, Anatol. 2008. Negative evidence and preemption: A constructional approach to ungrammaticality. *Cognitive Linguistics* 19(3). 513–531.

Stefanowitsch, Anatol. To appear. Collostructional analysis. In Graham Trousdale & Thomas Hoffmann (eds.), *The Oxford handbook of construction grammar*. Oxford: Oxford University Press.

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

Stefanowitsch, Anatol & Stefan Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1(1). 1–43.

Szmrecsanyi, Benedikt. 2006. *Morphosyntactic persistence in spoken English. A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin & New York: Mouton de Gruyter.

Thompson, Sandra A. & Yuka Koide. 1987. Iconicity and 'indirect objects' in English. *Journal of Pragmatics* 11(3). 309–406.

Tryk, H. Edward. 1986. Subjective scaling of word frequency. *The American Journal of Psychology* 81(2). 170–177.

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.

Zeldes. Amir. 2011. *Productivity in argument selection: A usage-based approach to lexical choice in syntactic slots*. Berlin: Humboldt University dissertation.

Author's address

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
Santa Barbara, CA 93106-3100
USA

stgries@linguistics.ucsb.edu