Collostructions: Investigating the interaction of words and constructions*

Anatol Stefanowitsch and Stefan Th. Gries
University of Bremen / University of Southern Denmark

This paper introduces an extension of collocational analysis that takes into account grammatical structure and is specifically geared to investigating the interaction of lexemes and the grammatical constructions associated with them. The method is framed in a construction-based approach to language, i.e. it assumes that grammar consists of signs (form-meaning pairs) and is thus not fundamentally different from the lexicon. The method is applied to linguistic expressions at various levels of abstraction (words, semi-fixed phrases, argument structures, tense, aspect and mood). The method has two main applications: first, to increase the adequacy of grammatical description by providing an objective way of identifying the meaning of a grammatical construction and determining the degree to which particular slots in it prefer or are restricted to a particular set of lexemes; second, to provide data for linguistic theory-building.

Keywords: construction, Construction Grammar, collocation, Fisher exact test, syntax-lexis interface

1. Introduction

In this paper, we develop and demonstrate an extension of collocational analysis specifically geared to investigating the interaction of lexemes and the grammatical structures associated with them. This method is based on an approach to language that has been emerging in various frameworks in recent years, and that does not draw a fundamental distinction between lexicon and syntax, but instead views all of language as consisting of linguistic signs.

Traditionally, the lexicon and the grammar of a language are viewed as qualitatively completely different phenomena, with the lexicon consisting of

International Journal of Corpus Linguistics 8:2 (2003), **209–243.**
ISSN 1384–6655/ E-ISSN 1569–9811

specific lexical items, and the grammar consisting of abstract syntactic rules. Various expression types that fall somewhere in between lexicon and grammar (i.e. various types of fully or partially fixed multi-word expressions) have been recognized but largely ignored (or at least relegated to the periphery) by mainstream syntactic theories (notably, the various manifestations of Chomskyan generative grammar).

The predominance of this view may be part of the reason why corpus linguists, until recently, have largely refrained from detailed investigations of many grammatical phenomena. The main focus of interest was on collocations, i.e. (purely linear) co-occurrence preferences and restrictions pertaining to specific lexical items. If syntax was studied systematically at all, it was studied in terms of colligations, i.e. linear co-occurrence preferences and restrictions holding between specific lexical items and the word-class of the items that precede or follow them.

More recently, however, the focus within corpus linguistics has shifted to a more holistic view of language. Several theories – for example, Hunston and Francis' *Pattern Grammar* and Lewis' theory of *lexical chunks* (Hunston & Francis 2000; Lewis 1993; cf. also Sinclair 1991; Barlow & Kemmer 1994) – have more or less explicitly drawn attention to the fact that grammar and lexicon are not fundamentally different, and that the long-ignored multi-word expressions serve as an important link between them.

In this respect, Pattern Grammar and Lexical-Chunk Theory are two relatively recent arrivals among a variety of approaches that have been emerging over the past two decades, and that share a view of both lexicon and (some or all of) grammar as consisting of linguistic signs, i.e. pairs of form and meaning – most notably the group of theories known as Construction Grammar, (e.g. Fillmore 1985, 1988; Kay & Fillmore 1999; Lakoff 1987; Goldberg 1995, 1999), but also Emergent Grammar (Hopper 1987; Bybee 1998), Cognitive Grammar (Langacker 1987, 1991), and at least some versions of LFG (cf. Pinker 1989) and HPSG (cf. Pollard & Sag 1994); note also that various approaches in ELT have advocated this insight more or less explicitly (cf. e.g. Pawley & Syder 1983). The meaningful grammatical structures that are seen to make up (most or all of) the grammar of a language are variously referred to by terms such as *constructions*, *signs, patterns, lexical/idiom chunks*, and a variety of other labels.

As we will show, this view of language makes the study of grammar more similar to the study of the lexicon, and it also makes it more amenable to investigation by corpus-linguistic methods. The method we propose has two main applications: first, to increase the adequacy of grammatical description, and

second, to provide data for linguistic theorizing and model-building. With regard to description, the method provides an objective approach to identifying the meaning of a grammatical construction and of determining the degree to which particular slots in a grammatical structure prefer, or are restricted to, a particular set or semantic class of lexical items. With regard to linguistic model-building the method provides data that may be used in investigating a variety of questions. The question we will mainly be concerned with is 'Are there significant associations between words and grammatical structure at all levels of abstractness?; other potential domains of application include diachrony, variation or (first or second) language acquisition.

This paper is structured as follows. Section 2.1 explicates the view that both lexicon and grammar are essentially repositories of meaningful units of various degrees of specificity. Section 2.2 introduces and justifies the methodology in some detail. Section 3 then sketches out how this methodology may be applied to successively more abstract grammatical phenomena, beginning with the verb *cause* with three different argument structures – *transitive*, *ditransitive*, and *prepositional dative* and moving on to a partially-fixed expression, [X *think nothing of* VP_{gerund}] (Section 3.1), to argument structures, specifically, the *into*-causative [S V O *into* VP_{gerund}] and the *ditransitive* [S V O_i O_d] (Section 3.2), and finally to even more abstract grammatical phenomena – progressive aspect, imperative mood, and past tense (Section 3.3).

2. Collostructional analysis

2.1 The theoretical background

While the method which we will develop below can yield insightful results for any of the frameworks mentioned in the introduction, we will – for the purposes of this paper – adopt the terminology and the basic assumptions of *Construction Grammar*, specifically, the version developed e.g. in Lakoff (1987) and Goldberg (1995). This theory sees the construction as the basic unit of linguistic organization, where *construction* is defined as follows:

> A construction is … a pairing of form with meaning/use such that some aspect of the form or some aspect of the meaning/use is not strictly predictable from the component parts or from other constructions already established to exist in the language (Goldberg 1996: 68; cf. also Goldberg 1995: 4 for a slightly more formal definition).

In other words, a construction is any linguistic expression, no matter how concrete or abstract, that is directly associated with a particular meaning or function, and whose form or meaning cannot be compositionally derived. The linguistic system is then viewed as a continuum of successively more abstract constructions, from words to fully-fixed expressions to variable idioms to partially filled constructions to abstract constructions.[1]

At the most specific end of the continuum are single morphemes (like [*mis*-V]/'wrongly, astray', [V-*ing*]/'act of', [N-*s*]/'plural') and mono-morphemic words (like *give* and *away*), followed by multi-morphemic words like *misgivings* or *giveaway*. We will retain the terms morpheme and lexeme for these (but they are sometimes referred to as morphological and lexical constructions). The definition also covers fully-fixed multi-word expressions (e.g. proverbial expressions like *He gives twice who gives quickly* or *Don't give up the day job*). Next, and slightly more abstract, there are fixed or variable multi-word-expressions including compounds (like *give-and-take*, or *care-giver*), phrasal verbs (like *to give up on sb*), lexically fully or partially filled idiomatic expressions (like *to give lip-service to sth* or [SUBJ *be given to* $N_{activity}$]/'X habitually does Y', as in *Linguists are given to making wild claims*). Finally, and crucially for the methodology we develop here, the definition also covers abstract syntactic structures like phrasal categories, argument structures, tense, aspect, mood, etc.

As an example of an abstract construction, take the English ditransitive subcategorization frame [S V O_i O_d], exemplified by *John gave Mary a book*. This subcategorization frame assigns a 'transfer' meaning (the notion that the referent of the subject transfers the referent of the direct object to the referent of the indirect object) to all expressions instantiating it, irrespective of the particular verb occurring in this frame. This is shown, for example, by the use of *hit* in *Pat hit Chris the ball*. *Hit* is a two-place verb whose meaning can roughly be glossed as '(some part of) X comes into forceful contact with (some part of) Y'. Clearly, nothing in its meaning points to a transfer of Y to some third participant. However, a sentence like *Pat hit Chris the ball* will consistently receive the interpretation 'Pat transferred the ball to Chris by coming into forceful contact with it' (cf. Goldberg 1995: 34–35). Since the syntactic configuration [S V O_i O_d] is directly associated with the meaning 'X transfer Y to Z', and hence with the semantic roles Agent, Recipient, and Theme, and since this meaning is not strictly predictable from its components or from other constructions of English, the ditransitive subcategorization frame must be seen as a construction.

Any actual utterance larger than a word is a simultaneous manifestation of several constructions. For example, the sentence *Pat hit Chris the ball* instantiates the subject-predicate construction (i.e. [SUBJ PRED]), the ditransitive construction just discussed, the past tense construction (i.e. [V-*ed*]/'past'), the noun-phrase construction, and the lexemes (or lexical constructions) corresponding to the individual words (cf. Goldberg 1996: 68).

Once words and the grammatical constructions they are associated with (for example, verbs and their argument structures) are seen as independent but meaningful units, the question arises which words can co-occur with which constructions. Put simply, the answer given by Construction Grammar is that a word may occur in a construction if it is *semantically compatible* with the meaning of the construction (or, more precisely, with the meaning assigned by the construction to the particular slot in which the word appears). For example, the verb *give* may occur in the ditransitive construction because verb and construction have the same meaning ('sb transfers sth to sb'). Note, however, that semantic compatibility does not have to mean semantic identity. For example, as just pointed out, the word *hit* does not have a transfer meaning; however, its meaning is compatible with a transfer meaning – hitting something may be a way of setting something in motion, which may serve as a *means* of transferring it to someone. Here, the ditransitive construction is said to *coerce* a transfer reading of *hit*. In such cases, a more abstract construction may add properties that are unspecified or underspecified in the more specific construction (such as a lexical item). For example, the verb *hit* only specifies an Agent (a Hitter) and a Theme (a Hittee). These are compatible with two of the roles specified by the ditransitive construction. Since *hit* does not specify a third role, this can be added by the ditransitive construction itself. With a semantically non-compatible word, this is not possible. For example, the verb *deprive* is not compatible with the meaning of the ditransitive construction: it is almost an antonym of it, and it specifies three roles that are not all compatible with those specified by the construction: an Agent (a Depriver) a Patient (a Deprivee), and a Theme (the Deprived Thing). Thus, **Pat deprived Chris the ball* sounds unacceptable (and is highly unlikely ever to occur in a corpus).

2.2 The methodology

The view of constructions introduced in the preceding section places particularly emphasis on the pairing of linguistic forms with linguistic meaning. In contrast, corpus linguistic approaches to language frequently focus on form (at

least in the initial stages of investigation). Corpus-based studies usually start from the (automatic or semi-automatic) collection of data from a corpus;[2] the treatment of semantic issues, for example in the areas of computer-aided lexicography and word-sense disambiguation, is typically based on a more-or-less-systematic interpretation of patterns emerging from a manual inspection of (i) a KWIC concordance display providing the node word in its context and/or (ii) the node word's collocates, i.e. frequent words within a user-specified span around the node word. An example of the former is Oh (2000), who analyzes the meaning differences between *actually* and *in fact* in American English; examples of the latter include Kennedy's (1991) investigation of the distributional characteristics of the semantically similar words *between* and *through* and Biber's (1993) collocate-based identification of word senses. The kind of collocational analysis exemplified by the latter two studies lends itself to a high degree of automatic preprocessing and has yielded many important insights, but it is extremely probabilistic with respect to grammatical structure. For the sake of computational ease, such analyses (tend to) disregard the grammatical structures in which a search word and its collocates occur and instead assume that sufficiently high raw frequencies of the collocates will sort out relevant results from accidental ones. Given the view of language introduced in Section 2.1 above, this approach is too imprecise. First, the more abstract constructions often do not contain any specific morphological or even lexical material that would allow the researcher to identify them in a traditional collocational analysis. Second, a given configuration of formal elements may represent more than one construction (for example, [V-*ed*] may represent the past-participle construction in addition to the past-tense construction for many verbs, and [S *be given to* N] may represent a simple passive use of *give*, as in *This diamond ring was given to Mary (by John)*, or it may represent the habituality-marking construction mentioned in Section 2.1, as in *John was given to generosity*). A traditional collocational analysis could never distinguish such cases.

In response to these shortcomings, we propose a type of collocational analysis which is sensitive not only to various levels of linguistic structure, but to the specific constructions found at these levels. We will refer to this method as *collostructional analysis*. 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);[3] crucially, such 'slots' can exist at different levels of linguistic structure (for example, the ditransitive construction may be said to have four slots corresponding to the subject, the verb, and the indirect and

direct objects, and the past-tense construction may be said to have a slot corresponding to the verb occurring in the past tense). Lexemes that are attracted to a particular construction are referred to as *collexemes* of this construction; conversely, a construction associated with a particular lexeme may be referred to as a *collostruct*; the combination of a collexeme and a collostruct will be referred to as a *collostruction*.[4]

Let us illustrate this methodology and the way it differs from traditional collocational analysis by means of the construction [N *waiting to happen*]. Table 1 gives a complete KWIC concordance of this construction from the British National Corpus 1.0 (BNC) sorted after L1. On the basis of such data, a standard concordancer will produce the collocate display shown in Table 2.

This kind of collocate list has a variety of obvious drawbacks which are all due to the fact that linear structure is at best a partial indicator of syntactic structure. Specifically, it implies that *business*, *horizon* and *company* occur in the N slot of this construction. However, as concordance lines 14, 24 and 28 in Table 1 show, this is not the case. Conversely, two words that do occur in this slot (*recovery* and *it* in lines 12 and 28 respectively) are not shown in Table 2 because they are at position L3. This is partly due to the fact that words like *just* may occur between N and *waiting to happen*, but, perhaps more importantly, it is also due to the fact that there are two syntactic realizations of the pattern, a noun post-modified by a participial clause (i.e. [$_{NP}$ *an* [$_{N'}$ [$_{N}$ *accident*] [$_{S}$ *waiting to happen*]]], cf. e.g. line 1) and a copular construction with N as the subject (i.e. [$_{S}$ [$_{NP}$ *an accident*] [$_{AuxP}$ *is*] [$_{VP}$ *waiting to happen*]]], cf. e.g. line 29). Thus, with a construction like this, it is not actually enough to pay attention to syntactic (tree) structure; instead, we need to analyze the construction at a more abstract level of syntactic representation, which could be informally represented as [$_{Head}$ N [$_{Modifier}$ *waiting to happen*]]. Extracting the lexemes occurring in the N slot under this definition requires item-by-item inspection and manual coding, but it guarantees an error-free list of collexemes for further analysis. We will present such a list shortly. Finally, note that *accident* and *disaster* occur in both the singular and the plural in Table 1; collostructional analysis collapses these into one figure for each corresponding lemma unless there is reason to believe that the construction is associated with only one particular word form.

Before we return to this construction, let us turn to the issue of attraction and repulsion and, thus, the issue of a suitable measure of association. Researchers have been interested in determining association strengths between word forms at least since Berry-Rogghe (1974), for example in the context of

Table 1. KWIC concordance for the *waiting to happen* construction (sorted after L1)

#	left context	node	right context
1	Stewart said that there was an accident	waiting to happen	and he feared lives would be lost.
2	the horse's knees. It was an accident	waiting to happen."	Recall stewards, dressed in day-glo bibs,
3	you had a cartoon about an accident	waiting to happen.	You could have saved the cartoonist's fee
4	Unless, of course, it was an accident	waiting to happen.	That insurer has 1,500 appointed
5	"Why?" "Because Stud's like an accident	waiting to happen,	that's why." "Oh, fuck off, Joey! I'm
6	the site say it was an accident	waiting to happen.	Video-Taped report follows JESSICA
7	the building means it was an accident	waiting to happen.	Unfortunately last night an accident did
8	the horse's knees. It was an accident	waiting to happen."	Blow for "blot on landscape" golf range
9	the return of his body. An accident	waiting to happen.	Charity stunt team warned you're playing
10	of it. Bands like that are accidents	waiting to happen	in a world where 99 per cent
11	actions which are little more than accidents	waiting to happen.	A little more patience and consideration on
12	yesterday: "I think the recovery has been	waiting to happen	for the last couple of months. It
13	Saturday was an accident that had been	waiting to happen.	I wrote to Sir Bob Reid, the
14	accident at the heart of the company	waiting to happen:	now IBM's signalling of the death of
15	not matter was the real constitutional crisis	waiting to happen,	vindication to all those Euro-sceptics who
16	which Coleman warned him of the "disaster	waiting to happen".	The identity papers seized by the FBI
17	– I'm pulling. "This is a disaster	waiting to happen,"	he added, in a prophecy that would
18	who said that it was "a disaster	waiting to happen".	Our hospitals are so short of cash
19	just had to be one monumental disaster	waiting to happen,	Leith later realised. But to start with,
20	marriage to Mandy Smith was a disaster	waiting to happen.	Urging Jagger to rebuild his marriage with
21	is a graphic example of a disaster	waiting to happen.	Over the weekend all attempts to salvage
22	one of these may be a disaster	waiting to happen.	In Lancashire towns like Oldham, Bolton
23	described in The Independent as" a disaster	waiting to happen".	The management of the economy has
24	– "Well – for a business disaster	waiting to happen,	you seem to have come off remarkably
25	develops this theme, identifying "disasters	waiting to happen"	associated with liquified natural gas, oil and
26	events of this week were an earthquake	waiting to happen.	Historians will argue over what was the
27	the first-half goal rush was an event	waiting to happen.	Young wingers are like young spin bowlers;
28	As if it [sex]'s just over the horizon,	waiting to happen	to me, as weird and wonderful as
29	residents are certain that "an accident is	waiting to happen".	Their fears – which focus on a
30	arguments that a new industrial revolution is	waiting to happen	in space are, for now, unconvincing. The
31	Cause" was a carefully planned invasion just	waiting to happen,	poised at the starting gate for the
32	and I can feel the dream just	waiting to happen,	gathering its energies from somewhere on
33	a graphic illustration of the disaster that's	waiting to happen	out there." Stuck fast: the Bettina Danica
34	in food production. A disaster was	waiting to happen.	Like so many cash crops, sugar is
35	that there may be many more Welkoms	waiting to happen,	and if racial conflict does spread in

identifying semantic differences between near synonyms (cf., e.g., Church & Hanks 1990). This strand of research has convincingly demonstrated that raw co-occurrence frequencies are not an ideal measure of association strength for both theoretical and empirical reasons: raw frequency counts do not take into account the overall frequencies of a given word in the corpus, and therefore the most frequent collocates of any given word are typically function words, which are often of little use, for example for the identification of subtle semantic differences between near-synonyms (cf. Manning & Schütze 2000: 153).

Table 2. Collocate frequencies for the [N *waiting to happen*] construction

L2		L1		R1		R2	
an	11	accident, disaster	9	in	3	the	2
a	6	accidents, been,	2	and, the, you	2	a, added, at,	1
the	3	is, just		a, associated,	1	could, fears, for,	
disaster	2	company, crisis,	1	blow, charity, for,		he, hospitals,	
accident, are,	1	disasters, earth-		gathering, he,		IBM's, identity,	
business,		quake, event,		historians, I,		if, insurer, its,	
constitutional,		horizon, that's,		Leith, like, now,		Jagger,	
dream, had, has,		was, Welkoms		our, out, over,		Lancashire, last,	
identifying,				poised, recall,		later, little,	
invasion,				that, that's, their,		management,	
monumental,				to, unfortunately,		me, report, seem,	
more,				urging,		so, space,	
revolution, than				video-taped,		stewards, stunt,	
				vindication,		there, to, why,	
				young		will, wingers,	
						with, wrote	

In a series of papers, Church and his collaborators address these problems and argue in favor of statistical, information-theoretical methods of quantifying (significant) degrees of association between words (i.e. degrees of collocational strength) (Church et al. 1990, 1991, 1994). However, while the basic argument is by now generally accepted, it is far from clear which method is optimally suited for linguistic research, and Church et al.'s work has triggered a number of studies proposing a variety of measures for this purpose (cf. Dunning (1993), Pedersen (1996); cf. Oakes (1998) as well as Manning & Schütze (2000) for overviews).

In principle, any of the measures proposed could be applied in the context of collostructional analysis, but most of them are problematic in at least one of the following ways: first, many of the proposed statistics involve distributional assumptions that are not justified: normal distribution and homogeneity of variances are just two such assumptions which are hardly ever met when dealing with natural language data, and which render suspicious any statistical method based on them (e.g. Berry-Rogghe's (1974) *z*-score, Church et al.'s (1991) *t*-score). Second, some statistics are particularly prone to strongly overestimating association strengths and/or underestimating the probability of error when extremely rare collocations are investigated (e.g. MI) – even proposed non-parametric improvements like the χ^2-statistic or Dunning's (1993)

log-likelihood coefficient still rely on the Chi-square distribution for significance testing and are, thus, unreliable given the kind of extremely sparse data frequently encountered in corpus-linguistic tasks (cf. Manning & Schütze (2000:175, Note 7), Weeber, Vos & Baayen (2000) and Gries (2003) for examples). As will become evident, the unreliability of these tests with respect to rare collocations is particularly problematic in the case of collostructions, since the vast majority of collexemes occurring within any given construction have a very low frequency in that construction (cf. Zipf's law).

There is one statistic that is not subject to such theoretical and/or distributional shortcomings, namely the Fisher exact test (cf. Pedersen 1996). It neither makes any distributional assumptions, nor does it require any particular sample size. Its only disadvantage is that a single test may require the summation of thousands of point probabilities, making it a computationally extremely intensive test procedure. Since precision is of the utmost importance in calculating collostruction strength, we will use the Fisher exact test in spite of its computational cost.

Like virtually all measures of collocation strength between two words w_1 and w_2, the Fisher exact test can be performed on a two-by-two table representing the single and joint frequencies of w_1 and w_2 (or in our case, between a construction and a potential collexeme) in the corpus.

Thus, to calculate the collostruction strength of a given collexeme L for a given construction C, we need four frequencies: 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 can then be entered in a 4-by-4 table and submitted to the Fisher exact test (or any other distributional statistic). Obviously, defining what counts as an instance of construction C may involve decisions on the part of the researcher that have to be justified on theoretical grounds.

To return to the [N *waiting to happen*] construction, consider Table 3, which represents the required frequencies for the noun *accident* (= L) and the [N *waiting to happen*] construction (= C) from the BNC. The figures in italics are derived directly from the corpus data, the remaining ones result from subtractions; 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*.

On the basis of this information, the Fisher exact test computes the probability of this distribution and all more extreme distributions (in the direction of H_1) with the same marginal frequencies. For the data in Table 3, the p-value

Table 3. Crosstabulation of *accident* and the [N *waiting to happen*] construction

	accident	**¬accident**	**Row totals**
[N *waiting to happen*]	*14*	21	*35*
¬[N *waiting to happen*]	8,606	10,197,659	10,206,265
Column totals	*8,620*	10,197,680	*10,206,300*

Table 4. Collexemes most strongly attracted to the [N *waiting to happen*] construction[8]

Collexeme (n)	$p_{\text{Fisher exact}}$ **(collostruction strength)**
accident (14)	2.12E-34
disaster (12)	1.36E-33
welkom (1)	4.46E-05
earthquake (1)	2.46E-03
invasion (1)	7.10E-03
recovery (1)	1.32E-02
revolution (1)	1.68E-02
crisis (1)	2.21E-02
dream (1)	2.45E-02
it (sex) (1)	2.83E-02
event (1)	6.92E-02

is 2.1216E-34,[5] indicating that, as would be expected, the association between *accident* and the [N *waiting to happen*] construction is very strong. The same computation can be performed for all other Ns occurring in this construction, and the Ns can then be ranked according to their strength of association (the Fisher exact p-values, that is) with the construction. This procedure results in Table 4.[6,7]

Although the main point of this analysis (as of the case studies presented below) is to exemplify the method, let us briefly point out some interesting aspects of our results. First, this construction is not typically found in dictionaries, the only exception being the *Collins Cobuild* family of dictionaries. This omission may be due to the fact that lexicographers perceived this construction as having no unique head noun under which to list it. Second, the one dictionary (or family of dictionaries) that does have an entry for it, Collins Cobuild, lists it under the head noun *accident*, which receives *a posteriori* support by the collostructional analysis (although collostructional analysis would suggest that it also be included under the head word *disaster*, where Collins Cobuild at least gives an example). Finally, Collins Cobuild gives the following definition.

> If you describe something or someone as **an accident waiting to happen**, you mean that they are likely to be a cause of danger in the future, for example because they are in poor condition or behave in an unpredictable way. (Collins Cobuild E-Dict. s.v. *accident*)

The negative connotation here is clearly due to the word *accident* rather than the construction. Note the absence of such negative connotations with the words *recovery* (line 12), *dream* (line 32), *it/sex* (line 28) and *event* (line 27) (Table 1). This would perhaps suggest that the construction should receive its own entry under *wait* with a more neutral definition along the lines of 'if you describe something as **waiting to happen**, you mean that it will almost certainly occur and that this is already obvious at the present point in time (often used with a negative connotation)'. The fact that *accident* and *disaster* are so strongly associated with the construction could be conveyed by an appropriate choice of examples.

3. Case studies

In this section, we will investigate a variety of constructions with respect to their most strongly attracted and repelled collexemes. The principal focus throughout this section is on the methodology itself; although we will provide some discussion of the results in each case, this discussion is aimed at pointing out the potential of the method rather than at providing detailed analyses of specific phenomena. The order of presentation approximately reflects the degree of abstractness of the constructions as discussed above. Unless otherwise noted, all case studies are based on the British component of the International Corpus of English (ICE-GB).

3.1 Words and variable idioms

3.1.1 *Cause*

We will begin with the analysis of a single word, the verb *cause*. As will presently become clear, collostructional analysis allows for a more fine-grained analysis than traditional collocational analysis even in the case of a single word.

Previous collocational analyses have shown that the verb *cause* collocates predominantly with words that have a negative connotation (i.e., that *cause*

predominantly has a 'negative semantic prosody', cf. e.g. Stubbs 1995). Some typical examples are shown in (1):

(1) a. There's a bone in my nose that's slightly bent and it's progressively **caused** slight breathing problems (ICE s1a-051 97)
b. Instead so Mill argued the only ground for making something illegal was that it **caused** harm to others (ICE s2b-029 106)
c. I am sorry to have **caused** you some inconvenience by misreading the subscription information (ICE w1b-026 115)

As these examples show, the negative prosody is due to the words that occur in the logical object position of *cause*. Table 5 shows the results of a collostructional analysis of the lexemes occurring in this position.

The results clearly confirm the claim that *cause* has a negative connotation. However, note that *cause* occurs in three different constructions: the transitive, as in (1a), the prepositional dative, as in (1b), and the ditransitive, as in (1c).[9] Using the collostructional method, we can go beyond the type of general analysis that is possible on the basis of Table 5, and look at the result arguments of each of these constructions separately (i.e. the objects of transitive and prepositional dative uses, and the second (or 'direct') objects of ditransitives, as well as the subjects of passives for each construction). The results of such a separate analysis are shown in Table 6.

Table 5. Collexemes of *cause* (all nouns encoding the result argument *of cause*)

Collexeme (n)	Collostruction strength	Collexeme (n)	Collostruction strength
problem (22)	9.03E-23	wear (2)	7.63E-05
damage (9)	1.86E-13	swelling (2)	1.92E-04
harm (5)	3.9E-11	concern (3)	2.7E-04
havoc (3)	1.24E-08	trouble (3)	4.64E-04
distress (3)	1.08E-07	collapse (2)	4.83E-04
inconvenience (3)	2.58E-07	disruption (2)	4.83E-04
cancer (4)	6.93E-07	casualty (2)	1.09E-03
injury (5)	1.25E-06	crack (2)	1.23E-03
injustice (3)	1.39E-06	acrimony (1)	1.46E-03
stampede (2)	6.39E-06	drowsiness (1)	1.46E-03
congestion (2)	1.28E-05	head-crash (1)	1.46E-03
extrusion (2)	1.28E-05	hiccough (1)	1.46E-03
stress (3)	2.51E-05	hyperinflation (1)	1.46E-03
change (6)	2.73E-05	neuropraxia (1)	1.46E-03
hardship (2)	4.46E-05	perplexity (1)	1.46E-03

Table 6. Collexemes of *cause* by construction

TRANSITIVE		PREPOSITIONAL DATIVE		DITRANSITIVE	
Collexemes	**Coll. strength**	**Collexemes**	**Coll. strength**	**Collexemes**	**Coll. strength**
problem (18)	3.30E-18	harm (3)	4.37E-10	distress (1)	4.54E-04
damage (7)	2.52E-10	damage (2)	5.47E-05	hardship (1)	4.54E-04
havoc (3)	8.74E-09	modification (1)	6.56E-04	discomfort (1)	5.19E-04
cancer (4)	4.39E-07	inconvenience (1)	8.43E-04	inconvenience (1)	5.84E-04
injury (5)	7.12E-07	famine (1)	9.37E-04	problem (2)	8.57E-04
injustice (3)	9.84E-07	delight (1)	1.59E-03	pain (1)	3.24E-03
stampede (2)	5.08E-06	problem (2)	1.83E-03	difficulty (1)	7.83E-03
congestion (2)	1.01E-05	disruption (1)	2.06E-03	night up (1)	1.89E-02
extrusion (2)	1.01E-05	accident (1)	1.66E-02		
change (6)	1.43E-05				

Clearly, *cause* has a 'negative prosody' in all three constructions; however, there are fundamental differences between the three constructions with respect to the exact type of negative result. The transitive construction occurs exclusively, and the prepositional dative predominantly, with external states and events; in contrast, the ditransitive construction encodes predominantly internal (mental) states and experiences.

The difference between the transitive and the ditransitive use of *cause* is intriguing, and has been missed by traditional collocational analyses. One reason for this difference may be found in the different argument structure of these two uses. In the transitive use, there are two participants – an Agent (the causer) and an (Effected) Patient (the result); in contrast, in the ditransitive there are three participants – an Agent (the causer) and a Theme (the result) that is (metaphorically) transferred to a Recipient; the metaphorical recipient of the result of an action is naturally interpreted as an experiencer of this result (see Section 3.2.2 below). This inclusion of an experiencer makes the ditransitive suitable for encoding mental states and experiences.

3.1.2 *The* [*X* think nothing of V_{gerund}] *construction*

Let us now move beyond the level of single words, beginning with a relatively concrete idiomatic expression, [X *think nothing of* V_{gerund}], exemplified in (2).

(2) a. In their present mood people would **think nothing of** mortgaging themselves for years ahead in order to acquire some trifling luxury like a jar of brandied peaches or a few leaves of tobacco. (BNC: EWF)

b. As a bachelor it seemed slightly shocking to Rupert that a colleague, even though an anthropologist, should **think nothing of** abandoning his wife when she was ill. (BNC: HA4)

We will be concerned with the verbs that appear in the V_{gerund} slot. This construction is found in many dictionaries; a typical definition is the following:

> If you think nothing of doing something that other people might consider difficult or strange, you consider it to be easy or normal, and you do it often or would be quite willing to do it (Collins Cobuild, s.v. *think*)

This definition makes clear that we are in fact dealing with a construction, as this meaning is not predictable from the component parts or other constructions of English; if we attempted to identify the meaning of this construction compositionally, we would expect it to mean something like 'to have a very low opinion of', in analogy to expressions like *think {the world/highly/not much/poorly/little} of* (and indeed this is a possible interpretation, although the OED is the only dictionary we are aware of which lists it).

Given a definition like the one cited, we would expect the construction to strongly attract verbs that refer inherently to undesirable and/or risky activities. However, it is not clear that there are many such verbs since what is undesirable or risky depends very much on context. Thus, this construction provides an extreme test for the collostructional method. Table 7 lists the results (from the BNC).

As might perhaps be expected given our concerns about the context dependence of the notions 'desirability' and 'riskiness,' there are no verbs that occur very frequently in this construction; also, note that there are no great

Table 7. Collexemes most strongly attracted to the [X *think nothing of* V_{gerund}] construction

Collexeme (n)	Collostruction strength	Collexeme (n)	Collostruction strength
haggle (1)	4.83E-04	beat (1)	2.74E-02
mortgage (1)	1.79E-03	check up (1)	3.38E-02
confide (1)	2.01E-03	eat (1)	3.92E-02
motor (1)	2.23E-03	stay (1)	5.36E-02
spend (2)	3.28E-03	walk (1)	7.45E-02
offer (2)	4.13E-03	hear (1)	1.17E-01
rip (1)	4.18E-03	take (2)	1.21E-01
leap (1)	6.02E-03	pay (1)	1.21E-01
hire (1)	7.50E-03	bring (1)	1.36E-01
wave (1)	9.78E-03	call (1)	1.54E-01
blow (1)	1.29E-02	get (2)	1.67E-01
abandon (1)	1.45E-02	go (2)	1.85E-01
hand (1)	1.70E-02	put (1)	2.09E-01
fly (1)	2.66E-02		

differences in the frequencies of the verbs that do occur in it. However, even under these circumstances, our measure of collostruction strength is able to rank the verbs; what is more, this ranking does indeed pick out a number of verbs denoting potentially risky activities (like *mortgage*, *confide*, *motor*, *leap* and *fly*) and verbs denoting potentially undesirable activities (like *haggle*, *rip*, *abandon* and *beat* – especially the first-ranked *haggle* seems to have a strongly negative connotation). Although one may not want to claim that the meaning of this construction could be deduced with a high degree of certainty from the list of verbs in Table 7, especially if taken individually, their prominence among the top collexemes clearly conveys a 'semantic prosody' that meshes well with the meaning of the construction. Incidentally, there are two lexemes identified by collostructional analysis as being repelled by the construction: the high frequency, low-content verbs *be* and *do*. Note that these would not help at all in identifying the meaning of the construction (for *be,* p = 7.52E-06; for *do*, p is only 0.469).

3.2 Partially filled and unfilled argument structure constructions

3.2.1 *The* into-*causative*

We will now turn to an argument-structure construction, albeit one that still includes a specific function word, [S_{agent} V $O_{patient/agent}$ *into*-$A^{gerund}{}_{resulting\text{-}action}$]. This construction, which we refer to as the *into*-causative, is exemplified in (3).

(3) a. He **tricked** me **into employing** him.
b. They were **forced into formulating** an opinion.
c. We **conned** a grown-up **into buying** the tickets.

In a brief discussion of this construction, Hunston and Francis (2000: 102–104, 106) impressionistically provide some raw frequency data concerning the verbs found in the V slot of the construction. On the basis of these data, they identify a strong tendency of the construction to occur with verbs denoting negative emotions (e.g. *frighten*, *intimidate*, *panic*, *scare*, *terrify, embarrass, shock, shame* etc.) or ways of speaking cleverly and deviously (e.g. *talk, coax*, *cajole*, *charm*, *browbeat* etc.). They propose that verbs entering into the *into*-causative usually (i) do not mean 'talk reasonably' and (ii) can also be used transitively; they go on to argue that both of the senses they have identified are associated with "some kind of forcefulness or even coercion" (Hunston & Francis 2000: 106). Before we present our own results, however, two aspects of Hunston and Francis' work are worth noting. First, although this construction has two slots for

Table 8. Collexemes most strongly attracted to the V slot of the *into*-causative

Collexeme	Collostruction strength	Collexeme	Collostruction strength
trick (92)	2.11E-267	delude (19)	8.83E-49
fool (77)	1.68E-187	talk (62)	2.38E-48
coerce (53)	1.15E-158	goad (18)	1.35E-46
force (101)	6.31E-136	shame (19)	1.28E-45
mislead (57)	9.57E-110	brainwash (13)	2.42E-37
bully (45)	2.53E-109	seduce (17)	2.56E-35
deceive (48)	5.94E-109	push (34)	6.66E-35
con (34)	4.41E-102	tempt (22)	3.37E-32
pressurise (39)	4.8E-101	manipulate (19)	3.3E-31
provoke (48)	4.05E-87	inveigle (10)	1.04E-30
pressure (30)	3.88E-85	hoodwink (10)	1.52E-29
cajole (28)	4.08E-85	panick (15)	7.75E-28
blackmail (25)	3.31E-64	lure (14)	1.23E-27
dupe (19)	7.77E-52	lull (11)	4.62E-26
coax (22)	6E-51	dragoon (8)	1.63E-25

verbs (V and A_{gerund}), Hunston and Francis confine themselves to a discussion of the V slot. Second, while Hunston and Francis comment on the notions 'force' or 'coercion' that at least one sense of the construction is associated with, the verbs *force* and *coerce* themselves are completely absent from their discussion and from the list of verbs they present.

Consider now Table 8, which shows the 30 verbs most strongly attracted to the V slot of the construction (data from the BNC).

Clearly, the results of the collostructional analysis differ strongly from the more impressionistic results presented by Hunston and Francis. First, the verb most strongly attracted to this construction is *trick*, whose collostruction strength is eighty orders of magnitude larger than that of the next-strongest collexeme, *fool*, or that of the most frequent verb in this construction, *force* (also note that second-ranked *fool* has a similar meaning to *trick*). Interestingly, neither of these verbs is mentioned by Hunston and Francis, nor do they fit the proposed semantic generalization ('negative emotions' or 'speaking cleverly'). Second, the verbs ranked third and fourth again share some semantic characteristics, namely those of 'force' and 'coercion' mentioned by Hunston and Francis. However, the collostructional analysis demonstrates that the construction is not only associated with the semantic notions 'force' or 'coercion' but also with the actual verbs *force* and *coerce*.

The data in Table 8 also show an interesting tendency: the collexemes appear to be ordered so that the very top of the list features verbs instantiating the two major sub-senses of the construction, namely 'trickery' (as exemplified by *trick/fool* as well as *mislead*, *deceive*, *con*, *dupe*, *delude*, *hoodwink* and *lull*) and 'force' (exemplified by *coerce/force* as well as *bully*, *pressurize*, *pressure*, *push* and *press-gang*). Intuitively less central senses of the *into*-causative appear much further down the list, for example:

– 'verbal coercion', instantiated by *blackmail* (as well as by *threaten*, which is not among the top thirty collexemes, but still a significant collexeme);
– 'positive persuasion', i.e. A's providing B with a positive stimulus in order to cause B to do something, instantiated by *cajole* and *coax*;
– 'negative persuasion', i.e. A's providing B with a negative stimulus in order to cause B to do something, instantiated by *goad* and *shame*.

Collostructional analysis has more to offer. While space does not permit an exhaustive characterization of the *into*-causative, note that the A_{gerund} slot of the construction can be subjected to the same kind of collostructional analysis; furthermore, it is possible to establish intra-constructional correlations between lexemes occurring in the V slot and lexemes occurring in the A slot. We will very briefly mention three interesting findings (cf. Gries & Stefanowitsch (Forthcoming)).

First, the most strongly attracted verb, *trick*, does not exhibit any semantic restrictions or preferences with respect to the (kinds of) A_{gerund} lexemes it co-occurs with frequently; these include

– action verbs (e.g. *do*, *give*, *work*);
– transfer verbs (e.g. *give*, *hand*);
– mental activity verbs (e.g. *believe*, *think*, *like*);
– perception verbs (e.g. *see*, *feel*);
– communication verbs (e.g. *tell*, *talk*, *say*).

Second, the A slots of other verbs of the same semantic group (that of 'trickery') are much more restricted: they prefer A_{gerund}s encoding mental activity or transfer, but generally disprefer action, perception, and communication verbs. Finally, the lexemes in the A_{gerund} slots of 'force' verbs exhibit a markedly different semantic tendency: the 'force' sense is mainly used with action verbs and transfer verbs, whereas communication verbs are rare and mental activity and perception verbs hardly occur at all.

In sum, collostructional analysis yields intriguing results: first, as before, it shows that there are associations between this construction and individual verbs, and that these are ranked in a way that lends itself to a meaningful interpretation; second, it allows us to expand on such an interpretation by potentially identifying the most strongly attracted gerunds as well as V-A_{gerund} correlations within the construction.

3.2.2 *The ditransitive*

Traditionally, ditransitivity is viewed as a verbal complementation pattern or subcategorization frame, i.e. as a purely syntactic property of individual verbs. In other words, it is assumed that verbs like *give*, *promise*, or *tell* are 'ditransitive verbs'; cf. the examples in (4a) to (4c):

(4) a. Mary gave John a book.
b. Chris promised Pat a car.
c. John told Mary a story.

If this view was correct, there would be no point in performing a collostructional analysis of ditransitivity, since it would result trivially in a frequency list of ditransitive verbs. However, there are several reasons for assuming that ditransitive syntax (i.e. [S V O_i O_d]) is a (meaningful) construction that exists independently of the specific verbs that occur in it. First, so-called 'ditransitive verbs' may also occur with other types of syntax (cf. e.g. *Mary gave freely to the poor* (intransitive prepositional), *Chris promised to be on time* (clausal complement), and *John told Mary of his adventures at sea* (transitive prepositional). Second, typical 'intransitive' verbs (like *blow*) or transitive verbs (like *throw*) may also occur with ditransitive syntax, as in *Mary blew John a kiss* or *Chris threw Pat the ball*), and if they do so, they receive an interpretation that is very similar to that of 'ditransitive' verbs. As mentioned in Section 2.1, the ditransitive construction can be represented in its active declarative form as [S_{agent} V $O_{recipient}$ O_{theme}].

It is crucial to the idea that all cases of ditransitive syntax instantiate a single argument-structure construction that such a construction may have a basic sense with several semantic extensions. In the case of the ditransitive, the basic sense is generally assumed to be 'X causes Y to have/receive Z' (cf. Goldberg 1995: 38; Pinker 1989: 73). Example (4a) instantiates this sense, while examples (4b) and (4c) instantiate extensions: the former is linked to the basic sense by virtue of the fact that the satisfaction conditions of the speech-act verb *promise* imply a transfer; the latter is a metaphorical extension based on the idea that

Table 9. The ditransitive construction: basic sense and extensions

Sense	Sample verbs
Basic sense:	
Agent causes recipient to receive theme	*give, pass, hand, …* *throw, kick, …* *bring, send, take, …*
Extensions on the basis of general semantic processes (Goldberg 1995: 38):	
A. *Satisfaction conditions imply that* agent causes recipient to receive theme	*guarantee, promise, owe, …*
B. *Agent enables* recipient to receive theme	*permit, allow, …*
C. Agent causes recipient *not* to receive theme	*refuse, deny, …*
D. *Agent acts to* cause recipient to receive theme *in the future*	*leave, bequeath, grant, …*
E. *Agent intends to* cause recipient to receive theme	*bake, make, build, …* *get, grab, earn, …*
Extensions on the basis of metaphor (Goldberg 1995: 147–150):	
F. Communication as transfer, e.g. *She told Joe a fairy tale.*	*tell, teach, fax, …*
G. Perceiving as receiving, e.g. *He showed Bob the view.*	*show, give a glimpse, …*
H. Directed action as transfer, e.g. *She blew him a kiss.*	*blow (a kiss), give (a wink), …*
Exceptions based on individual verbs (Goldberg 1995: 131–136):	*cost, charge, envy, forgive …*

communication is the exchange of objects (cf. Reddy 1979). The polysemy of the ditransitive construction has been most extensively discussed in Goldberg (1995); the extensions she posits are summarized in Table 9.

The results of the analysis are shown in Table 10.

Again, collostructional analysis demonstrates not only that there are associations between the ditransitive and specific verbs, and that these can be ranked, but it also yields results that bear on analyses of the ditransitive such as that presented by Goldberg.

The strongest collocate is *give*, which is clearly the verb most closely associated with the form and the meaning of the ditransitive construction, both in the minds of native speakers (cf. the informal experiment in Goldberg 1995: 35–36) and in the literature on the ditransitive. It is also, of course, the verb most similar in meaning to the ditransitive (the OED, for example, de-

Table 10. Collexemes most strongly attracted to the ditransitive construction

Collexeme	Collostruction strength	Collexeme	Collostruction strength
give (461)	0	allocate (4)	2.91E-06
tell (128)	1.6E-127	wish (9)	3.11E-06
send (64)	7.26E-68	accord (3)	8.15E-06
offer (43)	3.31E-49	pay (13)	2.34E-05
show (49)	2.23E-33	hand (5)	3.01E-05
cost (20)	1.12E-22	guarantee (4)	4.72E-05
teach (15)	4.32E-16	buy (9)	6.35E-05
award (7)	1.36E-11	assign (3)	2.61E-04
allow (18)	1.12E-10	charge (4)	3.02E-04
lend (7)	2.85E-09	cause (8)	5.56E-04
deny (8)	4.5E-09	ask (12)	6.28E-04
owe (6)	2.67E-08	afford (4)	1.08E-03
promise (7)	3.23E-08	cook (3)	3.34E-03
earn (7)	2.13E-07	spare (2)	3.5E-03
grant (5)	1.33E-06	drop (3)	2.16E-02

fines the relevant meaning using words like '*transfer*' and '*provide with*', which are clearly close paraphrases of '*cause to receive*'. It seems, then, that for the ditransitive, collostruction strength confirms the importance of semantic compatibility, and it also seems that strong collexemes of a construction provide a good indicator of its meaning (although the extreme polysemy of the ditransitive construction must be taken into account for a detailed analysis of both of these issues, a point to which we will return presently).

The list of significant collexemes also provides a crucial clue as to why some verbs are thought of as inherently ditransitive even though they also occur in other constructions, and why some verbs are not thought of as ditransitive even though they occur regularly in the ditransitive construction. Essentially, the stronger its collostruction strength with the ditransitive, the more likely a given verb is to be thought of as ditransitive. Most native speakers would agree that the first twenty verbs in Table 10 are felt to be ditransitive, but intuitions become considerably more varied below this point; the non-significant collexemes include mostly verbs that we would not think of as ditransitive.

Turning to the polysemy of the ditransitive, it is interesting to note that the basic 'transfer' sense is not overwhelmingly dominant in the list of the next most strongest collocates after *give*; in fact, it is only instantiated by four or five other verbs among the complete list of significant collocates: *send*, *award*, *lend*, *drop*, and perhaps *assign*. Instead, the next strongest collocates after *give* mainly

instantiate extended senses: eight of the nine extensions listed in Table 10 are instantiated by one or more of the fifteen strongest collocates; extension A by *offer*, *owe*, and *promise*, extension B by *allow*, extension C by *deny*, extension D by *grant*, extension E by *earn*, extension F by *tell* and *teach*, extension G by *show*, and the exceptional uses by *cost*. Thus, collostructional analysis may provide us with evidence for the high degree of polysemy of some constructions (such as the *into*-causative or the ditransitive) as compared to others (such as [N *waiting to happen*] or [*think nothing of* V_{gerund}]).

3.3 Tense/aspect/mood

3.3.1 *The progressive*

Let us now turn to even more abstract constructions, beginning with the progressive aspect. It is generally assumed that the progressive construction presents the action denoted by the verb as an ongoing process (cf., e.g., Jespersen 1931: 178; Dowty 1979: 145). It has also been noted that, as a consequence, verbs with a stative *aktionsart* (which inherently present a process as ongoing) do not generally occur in the progressive construction except under very specific circumstances (Lakoff 1970: 121).

From a corpus-based perspective, we would certainly not expect absolute restrictions on the ability of any verb to occur in the progressive aspect construction. However, it seems plausible that stative verbs will be infrequently instantiated among the most strongly attracted collexemes, but will make up a substantial proportion of the most strongly repelled collexemes.

Table 11 lists the 30 most strongly attracted and repelled collexemes. The results lend an overwhelming support to the traditional analysis. A full twenty of the 30 most strongly repelled collexemes are stative (namely all verbs except for *call*, *put*, *find*, *base*, *set*, *let*, *mention*, *get*, *marry*, *stop*); note especially that the ten most strongly repelled verbs are all stative.

In addition, a number of observations emerge regarding semantic verb classes. For example, motion/posture verbs (e.g. *go*, *sit*, *come*) as well as communication verbs (e.g. *talk*, *listen*, *speak*) are reasonably frequent among the most strongly attracted verbs, but are not instantiated at all among the most strongly repelled verbs. Also, among the stative verbs strongly repelled by the progressive, verbs denoting mental processes are particularly prominent.[10]

Table 11. Collexemes most strongly attracted to the progressive construction

attracted		repelled	
Collexeme (n)	**Collostruction strength**	**Collexeme (n)**	**Collostruction strength**
talk (234)	1.32E-94	be (448)	0
go (640)	1.08E-89	know (31)	1.01E-63
try (282)	8.86E-84	think (160)	4.05E-34
look (371)	4.41E-77	see (72)	6.36E-31
work (250)	2.14E-68	have (247)	1.93E-29
sit (100)	2.55E-57	want (44)	6.51E-21
wait (88)	6.17E-38	mean (15)	7.72E-17
do (539)	2.16E-36	need (5)	1.11E-14
use (264)	3.18E-29	seem (3)	1.02E-10
come (348)	9.65E-26	believe (11)	3.44E-09
run (113)	1.75E-25	call (30)	3.32E-08
move (104)	5.8E-19	put (93)	6.7E-08
live (101)	1.97E-17	remember (12)	9.49E-08
deal (57)	2.19E-16	find (56)	4.58E-07
walk (55)	9.34E-16	include (6)	2.76E-06
watch (46)	2E-15	agree (9)	4.45E-06
wear (48)	3.76E-14	base (2)	2.04E-05
write (123)	1.58E-13	set (34)	3.39E-05
listen (42)	2.18E-12	sound (6)	3.55E-04
seek (48)	8.66E-11	concern (3)	3.92E-04
fight (32)	2.63E-10	imagine (2)	4.97E-04
stand (57)	4.97E-10	let (10)	5.83E-04
study (31)	1.67E-09	mention (8)	1.04E-03
plan (28)	1.87E-09	exist (4)	1.13E-03
increase (54)	2.36E-09	get (294)	1.27E-03
sing (25)	3.54E-09	regard (2)	1.27E-03
approach (25)	5.13E-09	require (12)	1.3E-03
depend (43)	6.21E-09	marry (1)	1.86E-03
speak (71)	1.24E-08	stop (7)	2.13E-03
sell (38)	1.46E-08	indicate (3)	2.29E-03

3.3.2 *The imperative*

It is received wisdom that the imperative sentence type (or mood) serves a 'directive' function, more specifically, that of a request (at least in its 'direct' or 'prototypical' use). Characterizations of requests typically include the idea the speaker wants the hearer to perform the requested action, i.e. that it is desirable to the speaker (cf. Searle 1969: 66–67; Wierzbicka 1991: 205; Sadock 1994: 401). In addition, it is sometimes claimed that the imperative expresses

the speaker's assumption that the hearer will actually perform the requested action (cf. Wierzbicka 1991:205), or even that it places the hearer under an obligation to do so (cf. Sadock 1994:401), or that it presupposes a "power (authority) gradient" between speaker and hearer (Givón 1989:145).

We might, thus, minimally expect a prevalence of verbs encoding actions that yield results desirable from the point of someone else, i.e. the speaker; note that the verb most frequently used in the pragmatics literature to exemplify the imperative is *pass* (as in *Pass the salt!*). In addition, we might expect some reflex of the authority or obligation aspect of the imperative.

The data, however, tell a different story. Consider Table 12, which lists the 15 most strongly attracted collexemes of the imperative construction.[11]

Let us begin with the classes of verbs found to be strongly attracted to the imperative. Four of the verbs in Table 12 are clearly not action verbs in any sense (*see*, *worry*, *remember*, *note*). Furthermore, many of the action verbs that do occur are atypical in that they do not yield tangible results (*look*, *listen*, *hang on*, *check*, *try*, *keep*). While result-yielding action verbs do also occur, they are not nearly as dominant as might be expected (making up only a third of the top fifteen collexemes).[12]

Let us now turn to the issue of the desirability of the requested action: a cursory glance at Table 12 suggests that what is at issue is a result desirable

Table 12. Collexemes most strongly attracted to the imperative construction

Collexeme	Collostruction strength
let (86)	1.99E-97
see (171)	7.47E-80
look (74)	1.18E-24
listen (26)	4.05E-23
worry (21)	5.18E-22
fold (16)	9.25E-22
remember (35)	1.83E-18
check (21)	2.09E-17
process (15)	2.16E-17
try (47)	5.13E-17
hang on (17)	7.90E-17
tell (46)	1.30E-15
note (16)	2.96E-15
add (21)	2.64E-12
keep (28)	1.13E-11

from the point of the hearer rather than the speaker. This is confirmed by a closer look at the top ten verbs.

First-ranked *let* requires little discussion in this context. It occurs predominantly in the combination *let me*, as in example (5a) and rarely in other combinations as in (5b).

(5) a. **Let** me also point out what could happen to companies that don't innovate (ICE s2a-037 045)
b. **Let** the racket do the work with very little follow-through (ICE w2d-013 060)

Such examples could plausibly be omitted from the analysis on the grounds as those with *let's*; cf. Note 11. However, the basic fact, namely that *let* is used to encode situations that are portrayed as desirable to the hearer, holds for other verbs as well, specifically, for the verbs *see*, *look*, *listen* and *remember*, which are typically used as in examples (6) to (9).

(6) a. Just try it and **see** what happens (ICE s1b-002 064)
b. **See** also the section below on 'Students from abroad' (ICE w2d-003 049)

(7) a. **Look** what happened to Jimmy Carter (ICE s2b-021 012)
b. Just **look** at the beautiful scenery here (ICE s2a-016 037)

(8) Uhm <,> but then they said **listen** we need to you know <,> decide very promptly (ICE s1a-092 048)

(9) **Remember** that alcohol affects your judgment of both people and situations (ICE w2d-009 081)

Each of these verbs would merit its own discussion, but suffice it here to point out what they all seem to share (in addition to the hearer-desirability) is an attention-directing (or perhaps even discourse-organizational) function, the same can, of course, be said of *note* and *hang on*. Clearly, the requested actions are (portrayed as being) beneficial to the hearer rather than the speaker: the examples convey a sense of suggesting or advising rather than commanding or requesting (actually, these actions are also beneficial to the speaker, but not in the way typically associated with the imperative – rather, the requested actions serve to support the future cooperation and interaction between speaker and hearer in a way that is very similar to the use of *let me* exemplified in (5a) above). A very clear case of desirability to the hearer is also presented by fifth-ranked *worry*, which occurs exclusively in the phrase *don't worry*.

This leaves us with four more canonical imperatives, namely *fold*, *check*, *process*, and possibly *tell*. Of these, *fold* and *process* are typical result-yielding action verbs, but (i) as imperatives they both occur only in a single file of the corpus (cf. below Section 4) and (ii) any sense of beneficiality to the speaker is notably absent (cf. (10) and (11)). *Check* in (12) is result-yielding in some sense, but some of the examples also bear resemblance to the uses of *see*, *look* and *listen* exemplified above in (6) to (9).

(10) **Fold** the short edge to the centre (ICE w2d-019 044)

(11) **Process** until the mixture has formed a smooth purée (ICE w2d-020 137)

(12) a. **Check** it out (ICE s1a-033 186)
b. **Check** the condition of the drive belt periodically and replace it if it is excessively worn (ICE w2d-018 016)

Tell has some clearly directive uses, as in (13) but many uses are discourse-organizational (cf. (14)), and thus not unlike *see*, *look*, *listen*, and *note*.

(13) **Tell** him we are waiting for the order (ICE s1a-004 046)

(14) **Tell** us about Barcelona then (ICE s1a-046 422)

Although this analysis does not even begin to address the intriguing facts that collostructional analysis may ultimately reveal about the imperative, it clearly shows one thing: imperatives are apparently avoided with typical action verbs. This is doubtless due to the fact that such a use would be highly imposing. Instead, one major function of the imperative seems to be the organization of spoken or written discourse (of course, differences between the two registers may well exist).

To sum up, collostructional analysis has again picked out and ranked a number of verbs as significant collexemes of the construction in question, but, in contrast to the analysis of the progressive presented in the preceding section, the results do not straightforwardly support simple traditional analyses. Instead, the verbs picked out by collostruction strength provide evidence that one of the typical uses of the imperative is to direct attention in a low-imposition fashion.

3.3.3 *The past tense*

Before we conclude, we would like to emphasize that the applicability of collostructional analysis is not limited to the type of semantically relatively specific construction discussed so far. To drive home this point, let us briefly look at one

of the most abstract constructions of the English language: the past tense. Intuitively, there are no strong expectations, if any, that the past tense should be strongly associated with any particular verb at all. However, as Table 13 shows, there are both strongly attracted and strongly repelled collexemes even for this construction. For the top two collexemes, it is possible to come up with a partial motivation for this attraction: the attraction of *be* is at least in part due to its function as a passive marker (which – at least in the ICE-GB – is more frequent in the past tense, a fact that is in itself in need of explanation), while *say* is the verb standardly used in introducing direct and indirect speech in narratives (which are typically in the past tense for obvious reasons). Beyond this, we do not pretend to have even the beginning of a plausible explanation for the facts in Table 13 (although it does not seem impossible that such an explanation may ultimately be found); however the very fact that there are such relations of attraction and repulsions seems noteworthy enough to be reported, since it presents a huge problem for rule-based approaches to language.

Table 13. Collexemes most strongly attracted to the past tense construction

attracted		repelled	
Collexeme	**Collostruction strength**	**Collexeme**	**Collostruction strength**
be (6620)	0	know (159)	1.35E-26
say (1359)	1.81E-278	do (257)	7.23E-26
have (841)	1.1E-16	use (76)	3.01E-22
nod (19)	3.54E-14	put (106)	9.77E-19
die (57)	2.02E-12	get (339)	1.14E-15
become (150)	6.71E-12	see (184)	8.11E-15
tell (192)	8.86E-12	suppose (3)	1.18E-13
feel (152)	1.34E-11	saw (1)	4.84E-13
come (383)	1.13E-10	like (34)	1.22E-12
arrive (47)	4.08E-10	cut (10)	7.07E-12
start (90)	2.57E-08	work (49)	1.34E-11
decide (71)	2.94E-07	read (39)	3.16E-11
fall (54)	1.71E-06	talk (28)	3.98E-11
ring (34)	1.91E-06	remember (17)	7.8E-11
sit (47)	1.97E-06	hope (13)	3.62E-10

4. Conclusions

The collostructional analyses of a number of constructions have demonstrated several advantages of the method.

First, the descriptive adequacy of grammatical description is strongly increased. While simpler and more traditional collocate-based approaches already provide a huge improvement on purely intuitive analyses, we believe that collostructional analysis with its emphasis on (i) the grammatical structures in which collexemes are embedded and (ii) the quantification of the degree of attraction/repulsion has more precise results and more rewarding perspectives to offer, for example for lexicography and language pedagogy, to name just two fields of application where there are obvious practical advantages to knowing which lexical items are strongly associated with or repelled by a particular construction.

Second, the results presented above have implications for linguistic theorizing and model-building. Most importantly, the very fact that there are any dependencies at all between particular words and particular grammatical structures provides strong support for theories that view grammatical structures as signs, specifically for theories that view language as a repository of linguistic units of various degrees of specificity. If syntactic structures served as meaningless templates waiting for the insertion of lexical material, no significant associations between these templates and specific verbs would be expected in the first place (proponents of rule-based, open-choice theories could of course shift variable idioms out of core grammar to the lexicon, but this strategy would seem counterintuitive in the case of more abstract constructions, such as argument structure, tense, aspect, mood, etc.).

Finally, collostructional analysis in our view has implications for psycholinguistic studies of language acquisition. Goldberg suggests that the semantics of some of the most basic argument structure constructions (including the ditransitive) are identified by the child on the basis of the fact that a few flexible and semantically light verbs (e.g. *give* for the ditransitive) tend to account for the majority of the occurrences of these constructions in both input and output (Goldberg 1999; Goldberg et al. Forthcoming: 7–10). Goldberg et al. (Forthcoming: 11) provide initial evidence that

> it is the high frequency of particular verbs in particular constructions that allows children to note a correlation between the meaning of a particular verb in a constructional pattern and the pattern itself.

They emphasize the importance of token frequency with respect to (i) non-linguistic categorization and prototype formation and (ii) the identification of the semantic properties of novel constructions (they provide experimental support for the latter point, concluding that "high token frequency of a single general exemplar does indeed facilitate the acquisition of constructional meaning"; p. 13). We believe that collostruction strength is even more promising than raw frequency with respect to these issues. Since collostructional analysis goes beyond raw frequencies of occurrence, it identifies not only the expressions which are frequent in particular constructions' slots; rather, it computes the degree of association between the collexeme and the collostruction, 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. That is, collostructional analysis provides the analyst with those expressions which are highly characteristic of the construction's semantics and which, therefore, are also relevant to the learner.

Future research will have to refine and extend collostructional analysis in several ways. Extensions include, for example, a method for the analysis of distinctive collocates, which will enable the researcher to tease apart distributional and/or semantic differences between semantically similar constructions. Church et al. (1991) introduce a variant of the *t*–test as a measure of differences between near synonyms. The general logic of their procedure can be transferred to collostructional analysis, where it can serve to identify those collexemes that differentiate most strongly between two constructions. Gries and Stefanowitsch (in preparation) develop an appropriate extension of the methodology presented here applying it to various cases of grammatical alternations and choices.[13] Additionally, a systematic well-founded methodology for the investigation of intra-constructional correlations of the type mentioned in Section 3.2.1 needs to be developed (see Gries & Stefanowitsch (Forthcoming)). Finally, collostructional analysis takes the perspective of investigating the elements (e.g. verbs) occurring in particular slots within a construction. Reversing this perspective would mean to look at one particular verb to determine in which constructions it occurs significantly frequently. This would result in a statistically sound version of what Hanks (1996) referred to as a verb's behavioural profile.

On the computational level, the identification of important collexemes and, in fact, most collocate-based analyses, can be further improved by weighing all collexemes according to their degree of dispersion in the analyzed corpus (using, say, Carroll's D_2; cf., e.g., Oakes 1998 and Piao 2002). Consider the fol-

lowing example: the verb *process* occurs in the imperative 15 times, yielding a collostructional strength of 8.54E-17 while *hang on* occurs in the imperative 17 times, yielding a smaller collostructional strength of 3.66E-16. On the basis of collostructional strength, thus, *process* is more important for a subsequent interpretation. However, *hang on* occurs as an imperative within 12 corpus files (i.e., $D_2 = 0.36$) while *process* occurs as an imperative in a single corpus file only (i.e., $D_2 = 0$). Thus, one might in fact weigh *hang on*'s collostructional strength more heavily since the high collostructional strength of *process* to the imperative is only due to a single speaker/writer.

To conclude, we believe that collostructional analysis and its potential refinements open up many rewarding avenues of research in corpus linguistics as well as in syntactic theory, and we hope to stimulate further research in this area.

Notes

* The order of authors is arbitrary. We thank Thomas Berg and Adele Goldberg for their comments on earlier drafts of this paper.

1. Obviously, there are many differences between Construction Grammar and the other approaches mentioned in the introduction, and this definition glosses over many of these: most importantly, Cognitive Grammar does not include the idea of non-compositionality in its definition of a construction, and Pattern Grammar and ELT approaches typically require some lexical material to be present in an expression in order to count it as a lexical/idiom chunk or pattern.

2. We do not invoke the specific distinction here between corpus-driven and corpus-based studies; 'corpus-based' studies is to be understood in the general sense of the term.

3. For the moment, we will only consider as repelled items those which do occur, but occur less frequently than expected, although it would of course also be possible to include items that should have occurred on statistical grounds, but did not.

4. The technical terms *collostruction* and *collexeme* are obvious blends of the words *construction* and *lexeme* with *collocation*. Likewise, the term *collostruct* is derived from *collostruction* by analogy to the derivation of *collocate* from *collocation*.

5. All statistics reported in this paper were computed with the current version of the R package.

6. Table 3 is an instance where, strictly speaking, the application of the Chi-square test would have been possible. However, since the collostruction strengths of all lexemes occurring in the N slot and the [N *waiting to happen*] construction were ranked according to the p-values as explained above, it was necessary to compute them all in the same way so as to avoid different computational procedures influencing the ranking. Computationally

less demanding alternative to the Fisher exact test are the binomial approximation or Dunning's (1993) log-likelihood coefficient LL. Especially with large sample sizes, these yield very similar results (for many practical purposes at least).

One might nevertheless object to our ranking the lexemes occurring in the N slot according to the p-values obtained by the Fisher exact test since this would normally have to be done using effect sizes (like η^2 for ANOVAs, d for *t*–tests or r^2 for product-moment correlations; cf., e.g., Rietveld & van Hout 1993: 59). However, 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. For instance, in Table 3, 14 of the 35 occurrences of the [N *waiting to happen*] construction involved *accident* (i.e. 40%), yielding the p-value of 2.12E-34 mentioned above. If we had only observed 8 instances of *accident* in a total of 20 cases of the [N *waiting to happen*] construction in the same corpus with the same frequency of *accident* (i.e. again 40%), the p-value would accordingly be raised to 3.22E-20, indicating that this hypothetical collostruction is less noteworthy than the actually observed one. This sensitivity to frequency seems a desirable property for a measure of collostruction strength, given that 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). Finally, note that we will not place much emphasis on the question of whether a particular collostruction strength falls below standard levels of significance such as 0.05 or 0.01 – instead, we will mainly use the p-values as an indicator of relative importance of a collostruction (following, e.g., earlier work by Pedersen 1996; Pedersen et al. 2003).

7. It might be useful to return briefly to the weaknesses of traditional techniques of mere collocate analysis pointed out above in connection with Table 1 and Table 2 above. Without belaboring the obvious, note that the inclusion of the false hits *horizon*, *company* and *business* would distort the accurate results on the basis of manual coding considerably. *Horizon*, *company* and *business* result in p-values of 0.006, 0.059 and 0.127 respectively; in other words, merely using collocates would promote the false hit *horizon* to the fifth most strongly attracted lexeme in the N slot of the construction.

8. Given the low frequencies involved in this rare construction, no lexemes were found that are repelled by the construction (i.e. that did occur but significantly less frequently than expected). However, although it has sometimes been argued that such instances of repulsion will be fairly infrequent (cf., e.g., Church & Hanks 1990: 24; Church et al. 1991: 124), such lexemes are found for several of the constructions discussed below.

9. In addition, *cause* can occur as the matrix verb of a causative construction, as in *x caused y to do z*. However, this use is relatively infrequent, and it seems to us that the claims of a negative semantic prosody do not necessarily apply to it (Stubbs 1995 does not list any verbal collocates of *cause* that could be contributed by this use). We therefore ignore this use here.

10. The claim that communication verbs do not occur at all among the repelled collexemes is clearly too strong a statement. Note the verbs *call* and *agree*, which must be regarded as communication verbs in at least some of their uses (further examples among the strongly

repelled collexemes not listed here include *mention, guess, thank, express, acknowledge, reject, state, conclude, answer, accuse*). However, note that all of these are speech-act verbs (i.e. they convey an illocutionary force). As is well known, speech-act verbs are a systematic exception to the constraint that prevents non-stative verbs from occurring in the simple present without a habitual reading: they standardly occur in the simple present in performatives or performative-like utterances. Thus, they often appear in the simple present where all other non-stative verbs would require the progressive aspect (cf. Langacker 1991:251–252 for discussion). The fact that mental verbs are particularly prominent among the strongly repelled stative collexemes can be explained along similar lines (cf. Wierzbicka 1991:238 who analyzes such verbs as quasi-performative).

11. The strongly associated collexemes in Table 4 are based on a concordance of imperatives in the ICE-GB excluding hortative cases such as *Let's stop it for the moment* (ICE s1a-001 050). However, the results do not change substantially even if such hortative cases are included in the analysis.

12. In this connection note that the verb used most frequently in the literature to exemplify the imperative, *pass*, is only ranked 187th by the collostructional analysis.

13. Consider as a brief example the so-called 'dative alternation':

(i) a. *Mary gave John a book*. ditransitive (cf. Section 3.2.2)
b. *Mary gave a book to John*. prepositional dative

The results of our distinctive-collexeme analysis demonstrate that there is a variety of distinctive collexemes, i.e. collexemes that significantly distinguish between the constructions by significantly preferring one construction over the other. Consider (ii) and (iii) for just a few collexemes that are most clearly distinctive for the ditransitive and the prepositional dative respectively.

(ii) *give >>> tell >>> show >> offer > allow > cost >> teach >> buy, wish > earn > ask*

(iii) *put > bring > add > attach >> play > say >> limit > take >commit, confine*

Note that the collexemes distinctive for the ditransitive comprise several verbs of directed communication (e.g. *tell*, *offer*, *teach*, *ask*) whereas no such communication verb is distinctive for the prepositional dative. Also, while the distinctive collexemes of the ditransitive instantiate most of the constructional extensions listed above in Table 9, those of the prepositional dative comprise several verbs of caused-motion (e.g. *put*, *bring*, *attach*, *take*); this finding lends some support to the Construction Grammar analysis according to which the prepositional dative is analyzed as an instance of the caused-motion construction on independent grounds (see Gries & Stefanowitsch (in preparation) for more detailed discussion).

References

Barlow, M., & S. Kemmer (1994). A schema-based approach to grammatical description. In S. D. Lima, R. L. Corrigan, & G. K. Iverson (Eds.), *The reality of linguistic rules* (pp. 19–42). Amsterdam/Philadelphia: Benjamins.

Berry-Rogghe, G. L. M. (1974). Automatic identification of phrasal verbs. In J. L. Mitchell (Ed.), *Computers in the humanities* (pp. 16–26). Edinburgh: Edinburgh University Press.

Biber, D. (1993). Co-occurrence patterns among collocations: A tool for corpus-based lexical knowledge acquisition. *Computational Linguistics, 19*(3), 531–538.

Bybee, J. (1998). The emergent lexicon. In M. C. Gruber, D. Higgins, K. S. Olson, & T. Wysocki (Eds.), *Papers from the Thirty-Fourth Regional Meeting of the Chicago Linguistic Society* (pp. 421–435). Chicago: Chicago Linguistics Society.

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

Church, K. W., W. Gale, P. Hanks, & D. Hindle (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisition: Exploiting on-line resources to build up a lexicon* (pp. 115–164). Hillsdale, NJ: Lawrence Erlbaum.

Church, K. W., W. Gale, P. Hanks, D. Hindle, & R. Moon (1994). Lexical substitutability. In B. T. S. Atkins & A. Zampolli (Eds.), *Computational approaches to the lexicon* (pp. 153–177). Oxford: Oxford University Press.

Collins Cobuild E-Dict (1998). Glasgow: Harper Collins Publ.

Dowty, D. R. (1979). *Word meaning and Montague grammar. The Semantics of verbs and times in generative semantics and in Montague's PTQ.* Dordrecht: Reidel.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61–74.

Fillmore, C. J. (1985). Syntactic intrusions and the notion of grammatical construction. In M. Niepokuj, M. VanClay, V. Nikiforidou, & D. Feder (Eds.), *Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society* (pp. 73–86). University of California, Berkeley: Berkeley Linguistics Society.

Fillmore, C. J. (1988). The mechanisms of 'Construction Grammar'. In S. Axmaker, A. Jaisser, & H. Singmaster (Eds.), *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society* (pp. 35–55). University of California, Berkeley: Berkeley Linguistics Society.

Givón, T. (1989). *Mind, code, and context: Essays in pragmatics*. Hillsdale, NJ: Erlbaum.

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

Goldberg, A. E. (1996). Construction grammar. In K. Brown & J. Miller (Eds.), *Concise encyclopedia of syntactic theories* (pp. 68–71). Oxford: Pergamon.

Goldberg, A. E. (1999). The emergence of the semantics of argument structure constructions. In B. MacWhinney (Ed.), *The emergence of language* (pp. 197–212). Mahwah, NJ: Lawrence Erlbaum.

Goldberg, A. E., D. M. Casenhiser, & N. Sethuraman (Forthcoming). Learning argument structure generalizations. *Cognitive Linguistics*.

Gries, St. Th. (2003). Testing the sub-test: A collocational-overlap analysis of English *-ic* and *-ical* adjectives. *International Journal of Corpus Linguistics, 8*(1), 31–61.

Gries, St. Th., & A. Stefanowitsch (Forthcoming). Co-varying collexemes in the *into*-causative. In M. Achard & S. Kemmer (Eds.), *Language, culture, and mind.* Stanford, CA: CSLI Publications.

Gries, St. Th., & A. Stefanowitsch (In preparation). Extending collostructional analysis: A corpus-based perspective on 'alternations'.

Hanks, P. (1996). Contextual dependency and lexical sets. *International Journal of Corpus Linguistics, 1*(1), 75–98.

Hopper, P. (1987). Emergent grammar. In J. Aske, N. Beery, L. Michaelis, & H. Filip (Eds.), *Papers of the Thirteenth Annual Meeting of the Berkeley Linguistics Society* (pp. 139–157). University of California, Berkeley: Berkeley Linguistics Society.

Hunston, S., & G. Francis. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam/Philadelphia: John Benjamins.

Jackendoff, R. S. (1990). *Semantic structures*. Cambridge, MA: The MIT Press.

Jespersen, O. (1931). *A modern English grammar on historical principles*. Part IV: *Syntax*. London: Allen & Unwin.

Kay, P., & C. J. Fillmore (1999). Grammatical constructions and linguistic generalizations: The What's X Doing Y construction. *Language, 75*(1), 1–33.

Kennedy, G. (1991). *Between* and *through*: The company they keep and the functions they serve. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 95–110). London: Longman.

Lakoff, G. (1970). *Irregularity in syntax*. New York: Holt, Rinehart and Winston.

Lakoff, G. (1987). *Women, fire, and dangerous things*. Chicago: University of Chicago Press.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Vol. 1: Theoretical foundations*. Stanford, CA: Stanford University Press.

Langacker, R. W. (1991). *Foundations of cognitive grammar: Vol. 2: Descriptive application*. Stanford, CA: Stanford University Press.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago: The University of Chicago Press.

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. Hove, UK: Language Teaching Publications.

Manning, C. D., & H. Schütze (2000). *Foundations of statistical natural language processing*. 4th printing with corrections. Cambridge, MA: The MIT Press.

Oakes, M. P. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Oh, S.-Y. (2000). *Actually* and *in fact* in American English: A data-based analysis. *English Language and Linguistics, 4*(2), 243–268.

Oxford English Dictionary on CD-ROM. (1994). Version 1.15, 2nd edition. Oxford: Oxford University Press.

Pawley, A., & F. H. Syder (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richard & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–226). London: Longman.

Pedersen, T. (1996). Fishing for exactness. *Proceedings of the SCSUG 96 in Austin*, TX, 188–200.

Pedersen, T., S. Banerjee, & A. Purandare (2003). Ngram statistics package 0.53. http://www.d.umn.edu/~tpederse/code.html (downloaded 20 January 2003).

Piao, S. S. (2002). Word alignment in English-Chinese parallel corpora. *Literary and Linguistic Computing, 17*(2), 207–230.

Pinker, S. (1989). *Learnability and cognition. The acquisition of argument structure*. Cambridge, MA: MIT Press.

Pollard, C., & I. A. Sag (1994). *Head-driven phrase structure grammar*. Chicago/London: The University of Chicago Press.

Reddy, M. (1979). The conduit metaphor. In A. Ortony (Ed.), *Metaphor and thought* (pp. 284–324). Cambridge: Cambridge University Press.

Rietveld, T., & R. van Hout (1993). *Statistical techniques for the study of language and language behaviour.* Berlin/New York: Mouton de Gruyter.

Sadock, J. (1994). Toward a grammatically realistic typology of speech acts. In S. L. Tsohatzidis (Ed.), *Foundations of speech act theory. Philosophical and linguistic perspectives* (pp. 393–406). London: Routledge.

Sag, I., & Th. Wasow (1999). *Syntactic theory. A formal introduction*. Stanford: CSLI Publications.

Searle, J. (1969). *Speech acts*. Cambridge: Cambridge University Press.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stefanowitsch, A. (2001). Constructing causation: A construction grammar approach to analytic causatives. Doctoral dissertation, Rice University, Houston, TX.

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language, 2*(1), 23–55.

Weeber, M., R. Vos, & H. Baayen (2000). Extracting the lowest-frequency words: Pitfalls and possibilities. *Computational Linguistics, 26*(3), 301–317.

Wierzbicka, A. (1991). *Cross-cultural pragmatics. The semantics of human interaction*. Berlin/New York: Mouton de Gruyter.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.