43. Corpora and grammar

1. Introduction

The study of grammar is a relatively recent activity in corpus linguistics: for a long time, the word (more specifically, the orthographic word form) was the primary unit of investigation. As a consequence, the majority of corpus-linguistic studies have dealt with lexical issues (see article 58). The reason for this bias, which to some degree exists to this day, is mainly a methodological one: corpora are accessed via word forms, making them a natural choice for a focal point around which observations are made and theories are built. However, advances in automatic tagging and parsing (see article 13 and article 28) as well as the arrival of reasonably-sized corpora containing detailed manual or semi-manual grammatical annotation (see article 34) have increasingly enabled corpus linguists to shift their attention towards genuinely grammatical issues. This reorientation has not, for the most part, led researchers to discard the study of lexis. On the contrary, much of the recent quantitatively oriented corpus-based research of grammatical phenomena has been centrally concerned with the relationship between lexis and grammar. This article focuses on this line of investigation (see article 42 for more qualitative uses of corpora in the study of grammar).

2. Structure-sensitive collocates

A first, albeit indirect step toward the corpus-based investigation of grammar and its interaction with the lexicon is taken in a variant of collocational analysis that retrieves collocates on the basis of their part of speech and/or their syntactic relation to the node word rather than on the basis of their linear position. For example, a researcher may retrieve the adjectival collocates of a particular noun, the nominal collocates in subject position of a particular verb, etc. This method is primarily aimed at removing some of the noise of purely linear collocational techniques and thus achieves greater precision. As an example, consider Table 43.1, which shows the fifteen most frequent collocates directly preceding *time* (the most frequent noun in the BNC World corpus) as well as the fifteen most frequent *adjectival* collocates in the same position.

Retrieving only adjectival collocates removes many function words that are often not particularly informative with respect to the node word. This procedure is frequently used, and is implemented, for example, in the SARA concordancing tool (see article 33),

Tab. 43.1: Most frequent left collocates of *time* in a one-percent *n*-th line sample of the BNC World

All parts of speech		Adjectives only	
	F		F
the	266	*long*	38
first	104	*good*	11
this	96	*spare*	7
of	72	*little*	6
same	67	*present*	6
a	65	*whole*	5
that	49	*short*	5
in	44	*right*	4
some	39	*sufficient*	3
long	38	*best*	3
to	26	*appropriate*	3
any	25	*reasonable*	3
last	25	*real*	3
every	23	*different*	3
no	21	*particular*	3

which allows the user to restrict collocates to a particular part of speech. Clearly, grammar is still relatively peripheral in this approach, serving only to make lexical analyses more precise.

A related, but slightly more grammar-oriented approach involves choosing a grammatical frame and then investigating several or even all co-occurring words in this frame. As an example, Table 43.2 lists the 31 most frequent [ADJ + N] combinations in the

Tab. 43.2: Most frequent [A + N] combinations in an *n*-th line one-percent sample of the BNC

[ADJ + N] combination		[ADJ + N] combination	
	F		F
Prime Minister	102	*local government*	29
other hand	65	*European Community*	26
local authorities	44	*wide range*	26
long time	42	*working class*	25
Soviet Union	41	*armed forces*	24
other words	41	*old man*	23
local authority	37	*higher education*	23
labour party	37	*front door*	22
hon. friend	36	*social secu*rity	22
male speaker	35	*other thing*s	22
other side	34	*private sector*	21
other people	34	*other countrie*s	20
young people	33	*central govern*ment	20
chief executive	31	*great deal*	20
video-taped report	30	*recent year*s	20
hon. gentleman	29		

BNC World edition (note that here and throughout this article we use the format [POS + POS] rather than the authors' own representational formats).

Typically, such frames are relatively specific. For example, Justeson/Katz (1995a) use the frame [ADJ + N] for the purpose of disambiguating different senses of adjectives and Justeson/Katz (1995b) use the same frame as well as other NP frames (such as [N + N], [N + P + N], etc.) for the purposes of terminology identification; Krenn (2000) and Krenn/Evert (2001) use the frame [PP + V] to investigate the success of different association measures in identifying figurative expressions and support-verb constructions in German (see Evert/Kermes 2003 and Evert 2004 for a similar research question involving [ADJ + N] frames); Gries (2003) uses [ADJ + N] to distinguish between *-ic*/*-ical* adjective pairs; and Wulff (2003) uses the frame [ADJ + ADJ + N] to investigate factors influencing adjective order in English.

However, such frames can also be relatively abstract, as in Justeson/Katz (1991) who use $_{S}$[… ADJ … ADJ …] to identify degrees of association between antonymic adjectives or Stefanowitsch/Gries (2005), who use $_{S}$[… V … V …] to identify baseline co-occurrences of verbs within a sentence.

The majority of studies taking this approach retain a focus on words or lexically filled multi-word expressions and their properties. The inclusion of syntactic information is usually still aimed at improving the precision of collocational techniques rather than at the investigation of genuinely syntactic issues (although Krenn/Evert 2001 and Wulff 2003 are exceptions to some degree). However, by using grammatical frames to retrieve collocates, the crucial role of grammar is acknowledged and grammatical information is − implicitly or explicitly − taken into account in such studies. Investigating, for example, adjective-noun combinations does not only tell us something about the adjectives and nouns involved, but also about the syntax and semantics of the adjectival modification of nouns.

3. Collocational frameworks and grammar patterns

A second, much more explicit approach to the corpus-based investigation of grammar is the notion of collocational frameworks. These are defined as discontinuous sequences of two function words with an intervening content word such as [*a* + N + *of*] or [*too* + ADJ + *to*] by Renouf/Sinclair (1991), who show that words do not usually occur randomly in a given framework but typically belong to particular semantic classes associated with the framework in question. As an example, consider Table 43.3, which shows the twenty most frequent nouns occurring in the framework [*a* + N + *of*] in the BNC World.

Many of the nouns in this framework refer to quantities, and their ranking in this framework clearly does not reflect their ranking in the corpus as a whole. In other words, this − and other − collocational frameworks attract non-random, semantically restricted classes of words, a fact that Renouf/Sinclair interpret as evidence for the fact that such frameworks are linguistic units in their own right. More generally, they argue that the existence of such frameworks provides evidence for Sinclair's (1991) 'idiom principle' (although they do not use the term in this paper). The idiom principle states that "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into seg-

Tab. 43.3: Most frequent nouns in the framework [a + N + *of*] in an *n*-th line one-percent sample of the BNC World

Word		Word (contd.)	
	F		F
lot	143	*piece*	25
number	129	*kind*	25
couple	76	*range*	25
series	54	*sort*	20
bit	54	*sense*	20
member	46	*part*	20
result	44	*group*	20
variety	41	*total*	18
matter	38	*pair*	18
set	25	*form*	17

ments" (Sinclair 1991, 110); it contrasts with the open-choice principle, which states that "[a]t each point where a unit is completed (a word or a phrase or a clause), a large number of choices opens up and the only restraint is grammaticalness" (Sinclair 1991, 109). In Sinclair's view, both principles coexist, but the idiom principle is much more pervasive than traditionally assumed. Crucially in the present context, collocational frameworks demonstrate that the idiom principle is clearly not limited to what would traditionally be seen as idioms.

Collocational frameworks are probably the first attempt to genuinely investigate the relationship between grammar and lexicon on the basis of collocational methods. However, the strict definition of collocational frameworks as trigrams of the form [function word + content word + function word] limits the scope of this approach.

Partly in response to this limitation, Hunston/Francis (1999) develop the notion *grammar pattern*, which they define as "all the words and structures which are regularly associated with the word and which contribute to its meaning" (Hunston/Francis 1999, 37). They list three criteria that a structure must satisfy in order to be considered a pattern: first, it must be relatively frequent; second, it must be associated with a particular word or a semantic class of words (in other words, it must conform to some degree to the idiom principle discussed above); and, third, it must contribute a clearly identifiable meaning to expressions in which it occurs. These criteria cover not just collocational frameworks, but also a wide range of other structures, from partially lexically specified expressions like [V + possessive pronoun + *way* + PP/Adverbial] (as in *I made my way into the orchard*) or [V + NP + *into* + V-*ing*] (as in *They talked Lewis into becoming a Christian*) to fully abstract syntactic frames like [V + NP], [V + *that*] or [V + NP + NP].

As an example, consider Table 43.4, which lists all verbs occurring in the first slot of the pattern [V + *from* + V-*ing*] in the BNC World, where this pattern encodes the meaning 'the referent of the subject is a result of the process encoded by the gerund'.
This structure meets all three criteria for pattern-hood: it is limited to a small set of verbs that can be characterized semantically (they are motion verbs or change-of-state verbs), it contributes to the expressions in which it occurs (the verbs by themselves do not necessarily encode 'result' relations), and it is relatively frequent (it occurs 266 times in the BNC World). Note that a pattern is not just a formal string of words and gram-

Tab. 43.4: Verbs in the first slot of the pattern [V + *from* + V-*ing*] instantiating the meaning 'be a result of' in the BNC World

Word		Word (contd.)	
	F		F
come	151	*flow*	3
result	67	*grow*	2
arise	34	*originate*	2
stem	7	*ensue*	1
follow	5	*develop*	1
emerge	3		

matical categories: the string [V from V-*ing*] represents a number of distinct patterns in addition to the one in Table 43.4. For example, the most frequent pattern instantiated by this string is one encoding the meaning 'the referent of the subject actively does not engage in the activity encoded by the gerund'. This pattern is associated with a different set of verbs shown in Table 43.5.

Tab. 43.5: Nouns in the pattern [V + *from* + V-*ing*] instantiating the meaning 'actively not do something'

Word		Word (contd.)	
	F		F
refrain	248	*stop*	4
keep	21	*flinch*	3
abstain	18	*shy*	2
desist	17	*demur*	1
withdraw	5	*resist*	1

It is crucial to the notion of grammar patterns that not every string of words and/or grammatical categories counts as a pattern. For example, the string [V *in* NP] with an NP that encodes a location does not count as a pattern, since it is not restricted to a particular verb or class of verbs and contributes little or no semantic information to the verb it occurs with (Hunston/Francis 1999, 73).

The idea of grammar patterns is probably the most substantial advance in the corpus-based study of grammar in recent years. Still it suffers from a number of drawbacks, most importantly, first, a lack of quantification (it remains unclear how frequent a structure must be to count as a pattern); second, a lack of systematicity in its application (the criteria for pattern-hood are not always stringently applied); and, third, a lack of exhaustiveness (often, the most frequent verbs for a given pattern are left out of consideration completely).

4. Colligates

A third corpus-based approach to grammar is based on the notion *colligation*, introduced by Firth (e. g. 1968, 182) as a term for relations between grammatical categories. In corpus linguistics, the term is typically taken to refer to the co-occurrence of words

with particular grammatical categories (cf., e. g., Hoey 2000, 234). Although this notion has long been recognized in corpus linguistics, surprisingly little substantial work exists explicating and/or applying it. Where it is used, it is typically operationalized in terms of word classes occurring in a particular position relative to a node word, i. e. as collocation at the level of part-of-speech. For example, Table 43.6 lists the word classes occurring immediately to the left and to the right of the word *consequence* in a random sample of 100 concordance lines from the British National Corpus (note that *of* and the different forms of *have* have their own part-of-speech tag in the BNC).

Tab. 43.6: Word-class colligates of *consequence* (incl. punctuation) in a random sample from the BNC World edition

LEFT COLLIGATES		RIGHT COLLIGATES	
Word class	F	Word class	F
Indefinite Article	46	*of*	*55*
Adjective	24	Punctuation Mark	26
Preposition	19	Personal Pronoun	4
Cardinal Numeral	5	Adverb	3
Definite Article	4	Preposition	2
of	*2*	*have* (3Sg)	*2*
		Indefinite Article	1
		Conjunction	1
		Possessive Pronoun	1
		Definite Article	1
		Noun (Singular)	1
		Quotation Mark	1
		Verb (3Sg)	1
		Verb (Past Tense)	1

This interpretation of Firth's idea is very close to Renouf/Sinclair's concept of collocational frameworks: based on the lists in Table 43.6, we could, for example, hypothesize that the word occurs in the collocational framework [*a* + N + *of*] quite frequently (which it does: it occurs nine times in the sample used in Table 43.3 above). Consequently, this version of colligation analysis suffers from the same drawbacks. Most importantly, it does not take grammar into consideration beyond the level of word class and it retains a purely linear view of grammatical structure.

Recently, however, Hoey (e. g. 1997, 2004) has developed a considerably more comprehensive understanding of colligational relationships that remedies both of these shortcomings.

First, Hoey argues for a view of grammatical categories that goes beyond the notion of word class. He includes categories at considerably more abstract levels of grammatical structure, such as definiteness. As an example, take again the word *consequence*. In the sample also used above, this word occurs in indefinite contexts in 84 cases and in definite contexts in only 16 cases. Thus, Hoey would claim that *consequence* has a colligational preference for indefiniteness that goes beyond its preference for the determiner *a* at the position immediately to its left.

Second, Hoey includes hierarchical structure under his notion of colligation. Specifically, he suggests that the association of words to particular grammatical functions like subject, object, and complement can be insightfully investigated. For example, Hoey

observes that the noun *consequence* frequently occurs as part of a complement but is rarely found in object position. The sample used above confirms this observation (cf. Table 43.7). The word *consequence* colligates strongly with the grammatical function *adverbial*, followed by *subject complement* and *subject*. The function *object* is clearly avoided.

Tab. 43.7: Grammatical-function colligates of consequence in a random sample from the BNC World edition

Word class		Frequency
Subject		18
Subject of an equative construction	(17)	
Subject of a transitive construction	(1)	
Object		4
Object of a *have*-construction	(4)	
Adverbial		55
Adverbial with *as*	(25)	
Adverbial with *in*	(20)	
Adverbial with *of*	(9)	
Adverbial with *by*	(1)	
Complement		21
Subject Complement	(21)	
Apposition		2
Total		100

Hoey's notion of colligation is broad enough to include many studies of lexico-grammatical phenomena even where these do not use this term (cf. for example Mair on gerundial and infinitival complements after *begin* and *start* (Mair 2003) and on infinitival complementation in general (Mair 1990), Noël (2003) on infinitives, accusatives and *that*-clauses, and many similar studies).

Finally, it deserves mention that Hoey extends the idea of colligational associations to relationships between words and positions in texts (such as the beginning of a sentence, the beginning of a paragraph) etc. and between words and particular textual functions (such as disagreeing).

Hoey's view of colligation takes the crucial step towards a systematic corpus-based analysis of grammar and its relation to lexis. Like the work on collocational frameworks and grammar patterns, it shows that grammar and lexis are intertwined in intricate ways. However, also like this work it has so far not been given a strict quantitative underpinning − for example, it lacks a correction for expected baseline frequencies − and thus often remains impressionistic to some degree.

5. Collostructional analysis

The most recent attempt at a comprehensive framework for the corpus-based investigation of lexico-grammar is *collostructional analysis*, a set of methods for investigating the relationship between lexical items and (meaningful) grammatical structures based on their observed and expected co-occurrence in large corpora. Essentially, collostructional

analysis is an application of Firth's notions collocation and colligation within a framework that regards grammatical structure as consisting of meaningful signs, so-called 'constructions' (hence its name, a blend of *construction* and *collocational analysis*). Thus, collostructional analysis is rooted in a theoretical tradition that distinguishes it from other current approaches in the field. In addition, it is rooted in a methodological tradition that sets it apart from many implementations of earlier approaches.

As already mentioned, the theoretical tradition is one that assumes that (some or all) grammatical structures are best viewed as meaningful linguistic units (i. e. as signs in the Saussurean sense). There are a broad variety of theories that share this assumption (for example, Hunston/Francis' *Pattern Grammar*, see above); some of these theories differ radically from each other in many other respects, but collostructional analysis can usefully be applied within any of these. In fact, it could even be applied within frameworks that deny the possibility of meaningful grammatical structures altogether, as long as these theories posit *any* relationship at all between lexical items and grammatical structures (it would then become a version of colligational analysis, albeit a strictly quantified one, see below).

Collostructional analysis usually adopts the terminology and the background assumptions of one specific theory, Construction Grammar (Goldberg 1995). In this theory, a construction is any combination of linguistic entities whose formal or semantic properties are not fully predictable from its component parts and/or more general constructions ('rules') of the language. Crucially in the present context, constructions can have different degrees of specificity. Thus, the notion covers many of the structures referred to as 'collocational frameworks' or 'grammar patterns', but also the whole range of grammatical categories recognized by grammatical theory (including part-of-speech categories, grammatical relations, etc.). Thus, collostructional analysis captures most of the phenomena investigated in grammar-pattern analysis and colligational analysis in a unified methodological and theoretical framework.

Like these approaches, collostructional analysis has so far mainly focused on the relationship between lexis and grammatical structures. According to Construction Grammar, this relationship is determined by semantic compatibility: words occur in (slots provided by) a given construction if their meaning matches that of the construction. Collostructional analysis has confirmed this assumption from several perspectives.

The methodological tradition that collostructional analysis stems from is characterized on the one hand by a detailed, theoretically informed attention to different levels of linguistic structure, and on the other hand by the commitments of quantitative corpus linguistics: (i) the use of large, balanced corpora, (ii) the exhaustive retrieval of all instances of the phenomenon under investigation (even if this requires extensive manual post-editing), and (iii) strict statistical evaluation of the results.

5.1. Overview

If grammatical structures are linguistic signs on a par with lexical items, then the association between grammatical structures and lexical items (or other grammatical structures) can be investigated in the same way as associations between words. Instead of focusing on various types of relationships between two (or more) words, collostructional methods

focus on corresponding relationships between a construction and one or more words. There are currently three such methods, each with a different focus on the association between words and grammatical constructions:

- collexeme analysis is used in investigating the association between a construction and the words occurring in a particular slot in this construction (cf., e. g., Stefanowitsch/Gries 2003) – for example, between the verb *give* and the ditransitive construction as opposed to all other constructions;
- distinctive collexeme analysis is used in investigating the association between a word and (one member of) two or more semantically or functionally equivalent constructions (cf., e. g., Gries/Stefanowitsch 2004a) – for example, between the verb *give* and the ditransitive construction as opposed to the prepositional dative;
- covarying collexeme analysis is used in investigating the association between pairs of words occurring in two different slots in the same construction (cf., e. g., Gries/Stefanowitsch 2004b, Stefanowitsch/Gries 2005) – for example, the verb and the direct object in the ditransitive construction.

Like all collocation measures, the three types of collostructional analysis are best described in terms of a two-by-two distribution table like the one shown schematically in Table 43.8.

Tab. 43.8: Distribution table

	B	¬B	
A	O_{11}	O_{12}	R_1
¬A	O_{21}	O_{22}	R_2
	C_1	C_2	**N**

The three types of collostructional analysis differ only in terms of the values assigned to A, ¬A, B, and ¬B:

- for collexeme analysis, A corresponds to a given construction, ¬A corresponds to all other constructions in the corpus, B corresponds to a given word (lemma) occurring in a particular slot in A, and ¬B corresponds to all other words in occurring in the corpus;
- for distinctive collexeme analysis, A corresponds to one member of a pair of constructions, ¬A corresponds to the other member of the pair, B corresponds to a given word (lemma) occurring in a particular slot in A and/or ¬A, and ¬B corresponds to all other words occurring in A and/or ¬A;
- for covarying collexeme analysis, A corresponds to a particular word in Slot 1 of the construction, ¬A corresponds to all other words occurring in Slot 1, B corresponds to a particular word in Slot 2 of the construction, and ¬B corresponds to all other words occurring in Slot 2.

In principle, any distributional statistic can be applied to such a table (see article 36 and article 58) − clearly, nothing hinges theoretically on the choice of association measure. However, given the extremely asymmetric frequency distributions typically found in natural language data, it is highly desirable to use an exact test. In collostructional analysis,

the Fisher-Yates exact test is typically used. The association measure is then either the p-value or the negative base-10 logarithm of the p-value. This measure has the advantage of providing information about both the reliability of the obtained association/repulsion and its strength (cf. Stefanowitsch/Gries 2003, 238−239, n. 6 for detailed discussion), but alternative measures such as effect sizes, which would be independent of sample sizes could also be used (cf. Gries (to appear) for an example).

5.2. Collexeme analysis

Collexeme analysis is the most straightforward implementation of collocation analysis in a constructional framework: instead of a node word, the researcher retrieves all instances of a grammatical construction from the corpus, and instead of investigating collocates (i. e. words occurring in a user-defined span around the node word), one investigates the words occurring in a particular slot provided by that construction (such words are referred to as (potential) collexemes). The latter are typically lemmatized, but looking at word forms is equally possible and tends to yield results that are conceptually similar (cf. Gries (to appear)). Each word's frequencies are entered into a distribution table as described above, and the Fisher-Yates exact test (or some other appropriate statistic) is applied to these tables. As an example, consider the verb *give* and the ditransitive construction. The ICE-GB corpus (see article 20) contains 461 occurrences of *give* used ditransitively, 699 occurrences of other uses, 574 ditransitives with other verbs, and 136,930 uses that do not contain the verb *give*, and are not ditransitive. Table 43.9 shows this information in the appropriate form, together with the expected frequencies for each cell in parentheses.

Tab. 43.9: The distribution of *give* inside and outside of the ditransitive in the ICE-GB

	give	Other verbs	Row totals
Ditransitive	461 (9)	574 (1,026)	1,035
Other constructions	699 (1,151)	136,930 (136,478)	137,629
Column totals	1,160	137,504	138,664

Submitting these frequencies to the Fisher-Yates exact test yields a p-value of 0, indicating that the p-value is smaller than the smallest integer that home-issue computers will output (i. e., approx. 4.94e-324). Thus, the association between *give* and the ditransitive is an extremely significant one, but this in itself does not tell the researcher anything about the direction of the association, i. e. whether *give* is significantly more frequent than expected in the ditransitive (in which case it is referred to as a *significantly attracted collexeme*), or whether *give* is significantly less frequent than expected (in which case it is referred to as a *significantly repelled collexeme*). In order to determine this, the observed frequency must be compared to the expected one. In this case, there is a positive association, i. e. *give* is a strongly attracted collexeme of the ditransitive (in fact, the most strongly attracted one). One can now apply the same procedure to all 69 verbs occurring

in the ditransitive at least once, and rank the verbs in ascending order by their p-values. Table 43.10 shows the top twenty significantly attracted collexemes of the ditransitive, as well as the two only significantly repelled collexemes.

Tab. 43.10: Attracted and repelled collexemes in the ditransitive in the ICE-GB.

ATTRACTED COLLEXEMES		REPELLED COLLEXEMES	
Word	p	Word	p
give (461)	0	*make* (3)	2.72E-04
tell (128)	1.6E-127	*do* (10)	2.99E-03
send (64)	7.26E-68		
offer (43)	3.31E-49		
show (49)	2.23E-33		
cost (20)	1.12E-22		
teach (15)	4.32E-16		
award (7)	1.36E-11		
allow (18)	1.12E-10		
lend (7)	2.85E-09		
deny (8)	4.5E-09		
owe (6)	2.67E-08		
promise (7)	3.23E-08		
earn (7)	2.13E-07		
grant (5)	1.33E-06		
Allocate (4)	2.91E-06		
wish (9)	3.11E-06		
accord (3)	8.15E-06		
pay (13)	2.34E-05		
hand (5)	3.01E-05		

These results are typical for collexeme analysis in that they show two things. First, there are indeed significant associations between lexical items and grammatical structures. Second, these associations provide clear evidence for semantic coherence: the strongly attracted collexemes all involve a notion of 'transfer', either literally or metaphorically, which is the meaning typically posited for the ditransitive. This kind of result is typical enough to warrant a general claim that collostructional analysis can in fact be used to identify the meaning of a grammatical construction in the first place.

Concerning the repelled collexemes, there is little to say in this case, as there are only two such cases. However, it is worth noting that neither of these involves a notion of 'transfer' and thus they provide further evidence for semantic coherence. In this respect, the results in Table 43.10 are also typical; in many cases the number of repelled collexemes is much greater, and words displaying a lack of semantic coherence with respect to the construction are often predominant among these.

Note also that this method can easily be extended to words that do not occur at all in a given construction in a given corpus. For such words, collostructional analysis can determine whether they are significantly repelled by the construction or not. If they are significantly repelled, this may indicate that they are categorically barred from occurring in the construction in question − in other words, collostructional analysis allows the researcher to make principled statements about negative evidence (cf. Stefanowitsch 2006).

From a methodological perspective, a comment seems in order concerning the absence of post-hoc corrections in Table 43.10 (and in collostructional analysis in general): from a purely statistical perspective, it could be argued that the procedure described above constitutes a case of multiple testing, and thus the results would have to be corrected accordingly. This is not usually done in collostructional analysis for two reasons: first, there is a tradition in corpus linguistics to view each result as an independent test, and second, the values are mainly used for ranking items, and the ranking would not usually change due to post-hoc corrections.

5.3. Distinctive collexeme analysis

Distinctive collexeme analysis differs from collexeme analysis in that the association of a verb to a particular slot of a given construction is calculated not against its frequency in the corpus as a whole, but against its frequency in a corresponding slot in another specific construction or corresponding slots in several other constructions. This strategy is particularly useful for pairs of semantically, pragmatically, or otherwise functionally similar constructions (although it can, in principle, be applied to any pair or set of constructions). As an example, consider the famous pair consisting of the ditransitive construction and the prepositional dative. Many verbs can occur in both of these constructions, and this has led a number of researchers to posit a link between them. However, it is conceivable that some or all of these verbs have significant preferences towards one of the two. Take again the verb *give*, which was shown to be highly significantly associated with the ditransitive, but which also occurs in the prepositional dative. More precisely in the ICE-GB, it occurs in the prepositional dative 146 times, and there are 1,773 occurrences of the latter with other verbs; the frequencies for the ditransitive were already given above. Table 43.11 shows this information in the appropriate form (again with expected frequencies in parentheses).

Tab. 43.11: The distribution of *give* in the ditransitive and the prepositional dative in the ICE-GB

	give	Other verbs	Row totals
Ditransitive	461 (213)	574 (822)	1,035
To-dative	146 (394)	1,773 (1,525)	1,919
Column totals	607	2,347	2,954

Submitting these frequencies to the Fisher-Yates exact test yields a p-value of 1.835954E-120, which indicates that *give* highly significantly prefers the ditransitive even when compared to the prepositional dative. Since the comparison is only between these two constructions, this automatically entails that, of the verbs that occur in both constructions, *give* is the one least strongly associated with the prepositional dative. One can now apply the same procedure to all forty verbs that occur at least once in each of the two constructions in the ICE-GB, and rank the results for each construction in descending order of

Tab. 43.12: Distinctive collexemes in the ditransitive and the prepositional dative in the ICE-GB

Ditransitive (n = 1,035)		*To*-dative (n = 1,919)	
Word	p	Word	p
give (461:146)	1.84E-120	*bring* (7:82)	1.47E-09
tell (128:2)	8.77E-58	*play* (1:37)	1.46E-06
show (49:15)	8.32E-12	*take* (12:63)	2.00E-04
offer (43:15)	9.95E-10	*pass* (2:29)	2.00E-04
cost (20:1)	9.71E-09	*make* (3:23)	6.80E-03
teach (15:1)	1.49E-06	*sell* (1:14)	1.39E-02
wish (9:1)	5.00E-04	*do* (10:40)	1.51E-02
ask (12:4)	1.30E-03	*supply* (1:12)	2.91E-02
promise (7:1)	3.60E-03		
deny (8:3)	1.22E-02		
award (7:3)	2.60E-02		

the p-values. Table 43.12 shows the significantly distinctive collexemes for each construction.

These results are typical for distinctive-collexeme analysis in that they again provide clear evidence for associations between words and constructions and for semantic compatibility as the main principle governing these associations. Specifically, it has been argued by a number of authors that the ditransitive encodes a direct transfer of a theme from an agent to a recipient in a face-to-face situation, while the prepositional dative encodes a caused movement of a theme by an agent to a different location. The verbs in Table 43.12 reflect this distinction, most clearly in the case of the top collexemes *give* (direct transfer) and *bring* (motion to a different location).

Note that distinctive-collexeme analysis does not produce repelled collexemes, since the method assigns all collexemes to one or the other of the constructions under investigation and thus collexemes that are repelled by one construction are automatically attracted by the other.

The method has so far been applied to several classic cases of 'alternations' such as the verb-particle constructions or active vs. passive, but also pedagogically relevant alternations such as the *will* future vs. the *going-to* future or the *s*-genitive vs. the *of* construction. The extension to more than two alternative constructions referred to as multiple distinctive collexeme analysis can be used to investigate these cases and others in even more detail (cf. Gilquin (submitted)); straightforward extensions would be to investigate active vs. *be* passive vs. *get* passive or *will* future vs. *going-to* future vs. *shall* future etc.

5.4. Covarying-collexeme analysis

Covarying-collexeme analysis differs from the previous two methods in that it is not primarily concerned with the association between a word and a grammatical construction, but with the association between two words occupying specific slots in a given construction. Among the collostructional methods, it is thus most similar to traditional

collocate-based or colligate-based studies in that it focuses on the relationship between words, but it differs from these methods in that, unlike collocate-based studies, it takes grammatical structure into account, and unlike colligate-based studies, it defines the words via constructions rather than via word classes in a given span. The method is useful for investigating the kinds of issues that the more traditional methods address. Take again the ditransitive construction. This construction provides three slots in addition to the verb: an agent slot (the subject in an active-voice sentence), a recipient slot (the first object in an active sentence), and a theme slot (the second object in an active sentence). A strong collocational link is expected between the verb and the theme slot (since the theme is the thing undergoing the action denoted by the verb, selectional restrictions should hold). As an example consider the word *ask* and the theme *question*. *Ask* occurs 23 times in the ditransitive, 9 times with *question* and 14 times with other themes. *Question* occurs 9 times in the ditransitive, always with the verb *ask*. Given the total number of ditransitives in the corpus, all other figures can be derived automatically. They are shown in Table 43.13, together with the expected frequencies for each cell in parentheses.

Tab. 43.13: The distribution of *ask* and *question* in the ditransitive in the ICE-GB

	question	Other Object NPs	Row totals
ask	9 (0)	14 (23)	23
Other verbs	0 (9)	1,182 (1,173)	1,182
Column totals	9	1,196	1,205

Submitting these frequencies to the Fisher-Yates exact test yields a p-value of 5.7E-17, indicating a very strong association between these words in the ditransitive (they are referred to as a *significantly attracted collexeme pair*). One can now apply the same procedure to all verbs and all nouns in the theme slot and sort the results as before. Table 43.14 shows the top twenty significantly attracted collexeme pairs, as well as the only four significantly repelled ones.

These results are typical for item-based co-varying-collexeme analysis: first, like the other two methods, they provide clear evidence for associations between words and constructions; and second, they represent typical collocations based on semantic coherence between the words in question. In this case, this semantic coherence is anchored in frame-based knowledge about what people typically do with which object. This is obvious not only in the case of *ask a question*, but also in *offer s. o. a job*, *write (s. o.) a letter*, *send (s. o.) a cheque*, etc. In many cases the verb-theme combinations represent fixed or semi-fixed expressions (like *tell you what*, *take (s. o.) a minute*, *wish s. o. all the best*, or *drop (s. o.) a line*. Work on item-based co-varying collexemes has uncovered different types of semantic coherence between words in addition to the very concrete, frame-based one shown here, e. g. image-schematic coherence, coherence based on semantic prototypes, and coherence based on metaphors.

Tab. 43.14: Attracted and repelled pairs of co-varying V-Object collexemes in the ditransitive in the ICE-GB

ATTRACTED COLLEXEMES		REPELLED COLLEXEMES	
Word pair	p	Word pair	p
ask−question	5.70E-17	give−what	5.26E-08
tell−what	7.04E-15	give−that	0.0005
tell−that	1.51E-13	give−pound	0.0089
do−good	5.66E-09	give−one	0.0367
offer−job	7.21E-09	give−letter	0.0833
take−minute	2.82E-08	give−job	0.1385
write−letter	4.07E-08	give−money	0.2057
guarantee−place	4.99E-08	give−card	0.245
send−copy	2.12E-07	give−minute	0.245
wish−best	2.89E-07	give−it	0.3019
wish−success	2.89E-07	give−account	0.3789
tell−story	5.93E-07	give−love	0.3789
send−cheque	6.66E-07	give−quid	0.3789
set−deadline	4.14E-06	give−room	0.3789
take−hour	7.48E-06	give−freedom	0.4266
lend−money	1.61E-05	give−sth.	0.429
drop−line	2.48E-05	offer−what	0.5086
drop−note	2.48E-05	give−cash	0.564
tell−all about NP	3.67E-05	give−detail	0.564
tell−truth	3.67E-05	give−position	0.564

6. Outlook and desiderata

All of the methods discussed here − from collocational framework analysis and pattern grammar over colligation analysis to collostructional analysis − have produced a wealth of evidence concerning the association between lexical items and grammatical structures. However, there are several areas in which these methods can and should be improved.

6.1. Clustering collocates and collexemes

Collocation-based studies of words and/or grammatical categories and constructions are always faced with the problem that their result is simply a list of items ranked according to frequency or some statistical association measure. Such a list is not, in itself, an analysis of the phenomenon in question; typically, it must be grouped into semantic and/or syntactic classes before its relevance becomes clear. This grouping is usually done on the basis of intuitive common-sense criteria; clearly, a more objective, bottom-up approach would be highly desirable.

One statistical technique that lends itself well to this task is cluster analysis (see article 40), which has been used, for example, for the identification of syntactic categories (e. g. Brill et al. 1990), co-occurrence classes (e. g. Hindle 1990; Pereira/Tishby/Lee 1993; Li/

Abe 1996), and semantic classes (e. g. Waterman 1996; Schütze/Pedersen 1997; Schulte im Walde 2000).

In the context of collostructional analysis, for example, Gries/Stefanowitsch (to appear) use hierarchical agglomerative clustering to identify semantic classes among the collexemes in one slot by clustering them according to the collexemes in a different slot, for example, in the *way*-construction (as in *He made his way to the station*). They cluster the verb collexemes by the prepositional collexemes introducing the locative PP, and find clear and robust clusters of verbs of physical force (e. g., *force*, *push*), verbs of non-linear movement (e. g., *weave*, *wind*), and verbs of body-part related movement (e. g., *shoulder*, *elbow*). Such results provide strong evidence for the assumption that constructions are regularly associated with different related senses that are reflected by closely related groups of collexemes and, more generally, that the interpretation of collocates and collexemes can benefit greatly from further multivariate analysis.

6.2. The inclusion of additional variables

Collocation-based studies of lexis and/or grammar do not typically include additional variables, such as channel (spoken/written), register (formal/informal), dialect, gender, etc. However, a number of studies have shown that such variables have an influence, especially in the domain of lexico-grammar (cf., e. g., Biber/Conrad/Reppen 1998).

In order to allow for an easy integration of additional variables into collocation-based studies, Stefanowitsch/Gries (2005) propose an extension of the collocation-based method on the basis of configural frequency analysis using binomial tests (cf. von Eye 1990). This procedure allows the identification of positive and negative associations between variables in multi-dimensional contingency tables (as opposed to the two-by-two tables underlying traditional collocational methods). It may thus be used, among other things, to investigate triples of linguistic elements (for example, a construction and its collexemes in two different slots, as in Stefanowitsch/Gries' (2005) extension of co-varying-collexeme analysis, or two linguistic elements and an external variable such as channel, as in Gries (to appear), and Stefanowitsch/Gries (submitted)).

6.3. Word-sense sensitive analysis

For the most part, collocation-based studies of lexis and/or grammar ignore the fact that words are generally polysemous. This is mostly a matter of necessity, as there are currently no large corpora annotated for word senses (see article 26). However, it has been shown that the association of a given word to a construction may be contingent on specific senses of the word in question (cf. Roland/Jurafsky 2002). Clearly, thus, the inclusion of word senses into collocation-based approaches to grammar remains a highly desirable goal.

6.4. Dispersion

All approaches discussed here use co-occurrence frequencies to analyze the relationship between lexis and grammar, sometimes (but not always) subjected to statistical evalu-

ation. However, even where statistical procedures are used, these have so far failed to take into account the fact that high co-occurrence frequencies can be deceptive if they are due to the influence of a small number of corpus files or the output produced by a small number of speakers (cf. Gries 2006). Future work should therefore devise ways to weigh the frequency of co-occurrence of lexical and grammatical elements on the basis of their dispersion in the corpus as a whole.

7. Final remarks: Association measures vs. raw frequencies

Finally, it is still unclear whether association measures based on statistical tests are in fact superior to raw frequencies. On a priori grounds, statistical association measures may be argued to be superior due to their higher degree of sophistication, but this assumption has been called into question, for example, by Stubbs (1995) and Kilgarriff (2005). Ultimately, this is largely a matter of empirical research, which is still largely lacking. The experimental evidence that does exist, however, provides empirical support for the superiority of statistical association measures (more precisely, the Fisher-Yates exact test outlined above). Gries/Hampe/Schönefeld (2005, to appear) compare the predictive power of collostruction strength, frequency, and subcategorization probability by means of sentence-completion tasks and self-paced reading-time experiments and find that collostruction strength clearly outperforms the other variables. However, much more research is needed to confirm these results and provide solid evidence for more reliable generalizations in this exciting research area.

8. Literature

Biber, Douglas/Conrad, Susan/Reppen, Randi (1998), *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Brill, Eric/Magerman, David/Marcus, Mitch/Santorini, Beatrice (1990), Deducing Linguistic Structure from the Statistics of Large Corpora. In: *Proceedings of the DARPA Speech and Language Workshop*. Hidden Valley: PA, 275−282.

Evert, Stefan (2004), The Statistics of Word Cooccurrences: Word Pairs and Collocations. Unpublished PhD dissertation, University of Stuttgart.

Evert, Stefan/Kermes, Hannah (2003), Experiments on Candidate Data for Collocation Extraction. In: *Companion Volume to the Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*. Morristown, NJ: Association for Computational Linguistics, 83−86.

Eye, Alexander von (1990), *Introduction to Configural Frequency Analysis: The Search for Types and Antitypes in Crossclassifications*. Cambridge: Cambridge University Press.

Firth, John R. (1968), S*elected Papers of J.R. Firth 1952−59*. Edited by F. R. Palmer. London: Longman.

Gilquin, Gaëtanelle (submitted), Making Sense of Collostructional Analysis: On the Interplay between Verb Senses and Collexemes. Submitted to the electronic journal *Constructions*.

Goldberg, Adele E. (1995), *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Gries, Stefan Th. (2003), Testing the Sub-test: A Collocational-overlap Analysis of English *-ic* and *-ical* Adjectives. In: *International Journal of Corpus Linguistics* 8(1), 31−61.

Gries, Stefan Th. (2006), Some Proposals towards More Rigorous Corpus Linguistics. In: *Zeitschrift für Anglistik und Amerikanistik* 54(2), 191−202.

Gries, Stefan Th. (to appear), Corpus Data in Usage-based Linguistics: What's the Right Degree of Granularity for the Analysis of Argument Structure Constructions? In: Brda, Mario/Žic Fuchs, Milena (eds.), *Expanding Cognitive Linguistic Horizons*. Amsterdam/Philadelphia: John Benjamins.

Gries, Stefan Th./Hampe, Beate/Schönefeld, Doris (2005), Converging Evidence: Bringing Together Experimental and Corpus Data on the Association of Verbs and Constructions. In: *Cognitive Linguistics* 16(4), 635−676.

Gries, Stefan Th./Hampe, Beate/Schönefeld, Doris (to appear), Converging Evidence II: More on the Association of Verbs and Constructions.

Gries, Stefan Th./Stefanowitsch, Anatol (2004a), Extending Collostructional Analysis: A Corpus-based Perspective on 'Alternations'. In: *International Journal of Corpus Linguistics* 9(1), 97−129.

Gries, Stefan Th./Stefanowitsch, Anatol (2004b), Co-varying Collexemes in the Into-causative. In: Achard, Michel/Kemmer, Suzanne (eds.), *Language, Culture, and Mind*. Stanford, CA: CSLI, 225−236.

Gries, Stefan Th./Stefanowitsch, Anatol (to appear), Cluster Analysis and the Identification of Collexeme Classes. To appear in: Newman, J./Rice, S. (eds.), *Empirical and Experimental Methods in Cognitive/Functional Research*. Stanford: CSLI.

Hindle, Donald (1990), Noun Classification from Predicate Argument Structures. In: *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*. Pittsburgh, PA, 268−275.

Hoey, Michael (1997), From Concordance to Text Structure: New Uses for Computer Corpora. In: Lewandowska-Tomaszczyk, Barbara/Melia, James (eds.), *PALC'97: Practical Applications in Language Corpora*. Łódź: Łódź University Press, 2−23.

Hoey, Michael (2000), A World beyond Collocation: New Perspectives on Vocabulary Teaching. In: Lewis, Michael (ed.), *Teaching Collocations*. Hove, UK: Language Teaching Publications, 224−243.

Hoey, Michael (2004), Textual Colligation: A Special Kind of Lexical Priming. In: *Language and Computers* 49(1), 171−194.

Hunston, Susan/Francis, Gill (1999), *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam/Philadelphia: John Benjamins.

Justeson, John S./Katz, Slava M. (1991), Co-occurrences of Antonymous Adjectives and their Contexts. In: *Computational Linguistics* 17(1), 1−19.

Justeson, John S./Katz, Slava M. (1995a), Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. In: *Natural Language Engineering* 1, 9−27.

Justeson, John S./Katz, Slava M. (1995b), Principled Disambiguation: Discriminating Adjective Senses with Modified Nouns. In: *Computational Linguistics* 21(1), 1−27.

Kilgarriff, Adam (2005), Language is Never, Ever, Ever Random. In: *Corpus Linguistics and Linguistic Theory* 1(2), 263−276.

Krenn, Brigitte (2000), *The Usual Suspects: Data-oriented Models for the Identification and Representation of Lexical Collocations*. Saarbrücken: DFKI.

Krenn, Brigitte/Evert, Stefan (2001), Can we Do Better than Frequency? A Case Study on Extracting PP-verb Collocations. In: *Proceedings of the ACL Workshop on Collocations*. Toulouse, France, 39−46.

Li, Hang/Abe, Naoki (1996), Learning Dependencies between Case Frame Slots. In: *Proceedings of the 16th International Conference on Computational Linguistics*. Copenhagen, Denmark, 10−15.

Mair, Christian (1990), *Infinitival Complement Clauses in English*. Cambridge: Cambridge University Press.

Mair, Christian (2003), Gerundial Complements after *begin* and *start*: Grammatical and Sociolinguistic Factors, and How they Work against Each Other. In: Rohdenburg, Günter/Mondorf,

Britta (eds.), *Determinants of Grammatical Variation in English*. Berlin/New York: Mouton de Gruyter, 330–345.

Noël, Dirk (2003), Is There Semantics in All Syntax? The Case of Accusative and Infinitive Constructions vs. *That*-clauses. In: Rohdenburg, Günter/Mondorf, Britta (eds.), *Determinants of Grammatical Variation in English*. Berlin/New York: Mouton de Gruyter, 348–377.

Pereira, Fernando/Tishby, Naftali/Lee, Lillian (1993), Distributional Clustering of English Words. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. Columbus, OH, 183–190.

Renouf, Antoinette/Sinclair, John M. (1991), Collocational Frameworks in English. In: Aijmer, Karin/Altenberg, Bengt (eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman, 128–144.

Roland, Douglas/Jurafsky, Daniel (2002), Verb Sense and Subcategorization Probabilities. In: Merlo, Paola/Stevenson, Suzanne (eds.), *The Lexical Basis of Sentence Processing: Formal, Computational and Experimental Issues*. Amsterdam/Philadelphia: John Benjamins, 303–324.

Schulte im Walde, Sabine (2000), Clustering Verbs Semantically According to their Alternation Behaviour. In: *Proceedings of the 18th International Conference on Computational Linguistics*. Saarbrücken, Germany, 747–753.

Schütze, Hinrich/Pedersen, Jan O. (1997), A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval. In: *Information Processing and Management* 33, 307–318.

Sinclair, John M. (1991), *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stefanowitsch, Anatol (2006), Negative Evidence and the Raw Frequency Fallacy. In: *Corpus Linguistics and Linguistic Theory* 2(1), 61–77.

Stefanowitsch, Anatol/Gries, Stefan Th. (2003), Collostructions: Investigating the Interaction between Words and Constructions. In: *International Journal of Corpus Linguistics* 8(2), 209–243.

Stefanowitsch, Anatol/Gries, Stefan Th. (2005), Covarying Collexemes. In: *Corpus Linguistics and Linguistic Theory* 1(1), 1–43.

Stefanowitsch, Anatol/Gries, Stefan Th. (submitted), Register and Constructional Semantics: A Collostructional Case Study. Submitted to: Kristiansen, Gitte/Dirven, René (eds.), *Cognitive Sociolinguistics*. Berlin/Heidelberg/New York: Mouton de Gruyter.

Stubbs, Michael (1995), Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies. In: *Functions of Language* 2(1), 23–55.

Waterman, Scott A. (1996), Distinguished Usage. In: Boguraev, Branimir/Pustejovsky, James (eds.), *Corpus Processing for Lexical Acquisition*. Cambridge, MA: The MIT Press, 143–172.

Wulff, Stefanie (2003), A Multifactorial Corpus Analysis of Adjective Order in English. In: *International Journal of Corpus Linguistics* 8(2), 245–282.

Anatol Stefanowitsch, Bremen (Germany)
and Stefan Th. Gries, Santa Barbara, CA (USA)