Extending collostructional analysis

A corpus-based perspective on 'alternations'*

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

This paper introduces an extension of distinctive-collocate analysis that takes into account grammatical structure and is specifically geared to investigating pairs of semantically similar grammatical constructions and the lexemes that occur in them. The method, referred to as 'distinctive-collexeme analysis', identifies lexemes that exhibit a strong preference for one member of the pair as opposed to the other, and thus makes it possible to identify subtle distributional differences between the members of such a pair. The method can be applied in the context of what is sometimes referred to as 'grammatical alternation' (e.g. the dative alternation), but it can also be applied to other choices provided by the grammar (such as the two future tense constructions in English). The method has two main applications. First, it can reveal subtle differences between seemingly synonymous constructions, many of which are difficult to identify on the basis of more traditional approaches. Second, it can be used to investigate the very notion of 'alternation'; we show that many alternations are much more restricted than has hitherto been assumed, and thus confirm the claims of recent, non-derivational views of grammar.

Keywords: (diathesis) alternation, construction, collocation, collostruction, distinctive collexemes, syntactic variation, syntax-lexis interface, Fisher exact test

1. Introduction

Pairs of semantically more-or-less equivalent expressions, like those shown in (1)–(5), have captured the attention and imagination of researchers working in many different theoretical paradigms:

(1) a. John sent Mary the book.

International Journal of Corpus Linguistics 9:1 (2004), **97–129.**
ISSN 1384–6655/ E-ISSN 1569–9811

b. John sent the book to Mary.

(2) a. Picasso painted this picture.
b. This picture was painted by Picasso.

(3) a. John picked up the book.
b. John picked the book up.

(4) a. the university's budget
b. the budget of the university

(5) a. John will send Mary a book.
b. John is going to send Mary a book.

Some of these pairs, like the ones in (1)–(3) (and to some degree (4)) have played, and continue to play, an important role in the development of both formal and functional theories of language structure, while others, like (4) and (5), have a firm place among those areas of English grammar that are deemed worthy of detailed discussion in learner grammars of English. Pairs like those in (1)–(4), which we will refer to as *alternating pairs*, are typically discussed in terms of the formal, functional and/or semantic similarities and differences between their members.

Early transformational grammar and most of its modern descendants are concerned with the formal relationship between the members of such a pair, including sometimes the shared semantic restrictions on corresponding slots. This relationship is typically captured in terms of some derivational mechanism (such as a transformational rule) relating both members of a pair to the same underlying structure (cf. Chomsky 1957; Lees 1960 for early examples).

In contrast, discourse-functional approaches have focused on functional differences between the two members, notably, the different ways in which they package information flow. Such approaches typically postulate general linear precedence principles concerning notions like topicality, thematicity, givenness, or animacy, and set out to show that speakers choose one or the other variant depending on which one allows them to package the content that they want to express in accordance with these principles (cf. e.g. Givón 1993: esp. Ch. 11). The formal properties of the alternating pair in question are usually not addressed explicitly in these approaches, nor is the nature of the relationship between its two members.

More modern generative approaches, while still interested in the formal properties of these alternating pairs, focus mainly on the question which of the two members of a pair represents the 'basic argument structure' of a given

verb (or, in the case of the alternating pair in (4), a given head noun) and the question what semantic conditions determine whether a given word can undergo the change in argument structure that is necessary for it to occur in the 'non-basic' member (such changes are sometimes described in terms of 'lexical rules') (cf. e.g. Pinker (1989), Levin (1993), see also Fisher, Gleitman & Gleitman (1991)).

Semantic considerations also play a fundamental role in recent construction-based approaches to language, which have dispensed completely with the notion of positing one of the members of an alternating pair as basic and the other as derived, or deriving both from the same underlying source. Grammar is seen as a repository of linguistic signs (i.e. pairings of form (in this case, grammatical structures) and meaning) which are referred to as 'constructions' (Fillmore (1985), Lakoff (1987), Langacker (1987), Goldberg (1995), cf. also Stefanowitsch & Gries (2003)). In these approaches, each member of an apparent 'alternation' is a construction in its own right (Langacker 1991: Section 11.1 and Goldberg 2002); they appear to 'alternate' only because they are in a partial paraphrase relationship when used with certain lexical items (see esp. Goldberg (2002), who argues forcefully for this position, and shows that a construction containing a given verb shares more semantic and syntactic properties with constructions of the same kind containing different verbs than with the other construction of the alternating pair containing the same verb). In this approach, there is no fundamental difference between pairs like those in (1)–(4) and that in (5), which is why we will extend the term *alternating pair* to the latter.

In construction-based approaches, the question whether a given verb/noun may occur in one or both members of an alternating pair is seen in terms of semantic compatibility. A word may occur in a given construction if its meaning is compatible with the meaning of the construction; it may 'alternate' between two constructions if (or to the degree that) the word's meaning is compatible with the meanings of both constructions. In the context of alternating pairs, a focus on constructional semantics and semantic compatibility raises several questions: first, what exactly are the (often seemingly tenuous) semantic differences between the members of such a pair; second, how productive is the 'alternation' in actual usage, i.e. which verbs/nouns occur freely in both constructions, and which have strong biases towards one of them; and third, is a constructional, non-derivative approach plausible given the answers to the first two questions. We believe that a method that extends the notion of distinctive collocates in the context of our previously proposed 'collostructional analysis'

may provide answers to these questions. Section 2 will present this method, and Section 3 will apply it to the alternating pairs in (1)–(5) above. Section 4 will discuss some fundamental methodological issues.

2. Distinctive collexeme analysis

It is widely agreed in corpus linguistics that the linguistic context of a given word (or phrase) holds important clues to its semantic and syntactic properties, and thus the analysis of this context is a fundamental aspect of corpus-linguistic research. One way of analyzing a word's typical contexts is by extracting its collocates, where collocates are simply defined as 'words that occur (with a frequency that is significantly above chance-level) in a given span around the node word'. The span may vary in size; for example, it is ±1 in Kennedy's (1991) investigation of *between* and *through*, ±3 in Stubbs' (1995) semantic analysis of the verb *cause*, and ±5 words in Church and Hanks' (1990) analysis of *doctor*. In other words, the co-occurrence of words is investigated from a purely linear perspective on language that deliberately disregards syntactic structure. The studies in question assume that the noise created by this strategy can be filtered out by statistical methods, or even that relevant collocates will simply outnumber non-relevant ones. These assumptions have provided corpus linguists with intriguing data that have often led to interesting results, but the deliberate disregard of syntactic structure limits the usefulness of the method for fine-grained syntactic description (apart from limiting its credibility with syntactic theorists). In particular, the method is of limited use in the analysis of the relationship between lexical items and individual constructions or alternating pairs such as those discussed above. In recent work, we have combined collocate analysis with a close attention to the syntactic and semantic structures in which words occur in order to arrive at noise-free data that allow a much more fine-grained analysis of the dependencies and interactions between single words and grammatical constructions (cf. Stefanowitsch & Gries 2003), or between different words in a given grammatical construction (cf. Gries & Stefanowitsch, forthcoming).

This procedure, which we refer to as *collostructional analysis* (a blend of *construction* and *collocational*), can be extended to the analysis of alternating pairs of constructions and their relative preferences for words that can (or should be able to) occur in both of them. The logic behind this extension is basically that of Church et al.'s (1991) distinctive collocate analysis, which uses a

variant of the *t*-test as a measure of dissimilarity of semantically similar words on the basis of their lexical collocates (for example, Church et al. show how their *t*-test can identify collocates that distinguish between the adjectives *strong* and *powerful*).

We propose a similar method for the analysis of alternating pairs, differing from Church et al. in that we look at near-synonymous (or functionally near-equivalent) constructions rather than words, and that we focus on words appearing in particular slots in these constructions rather than at all words within a given span (we refer to such words as *collexemes* of the construction(s) in question).[1]

A variety of measures has been proposed to determine association strengths (for example, the *t*-test just mentioned as well as Berry Rogghe's (1974) *z*-score, Church & Hanks's (1990) pointwise mutual information or Dunning's (1993) log-likelihood coefficient). In principle, any of these could be applied in the context of distinctive collexeme analysis, but they are problematic in several ways. First, they often involve distributional assumptions – specifically, normal distribution and homogeneity of variances – that are hardly ever justified when dealing with natural language data (e.g. the *z*-score and the *t*-score). Second, they tend to strongly overestimate association strengths and/or underestimate the probability of error when extremely rare collocations are investigated (esp. MI). Even the non-parametric log-likelihood coefficient proposed by Dunning (1993), which is an improvement over parametric statistics, relies on the Chi-square distribution for significance testing and is, 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 (2003c) for examples). This unreliability with respect to rare collocations is particularly problematic in the case of collostructions, since the vast majority of collexemes occurring in any given construction have a very low frequency in that construction (cf. Stefanowitsch & Gries (2003)).

In order to avoid these problems, we will use the Fisher exact test (cf. Pedersen 1996), which is not subject to such theoretical and/or distributional shortcomings. As an exact test, it does not make any distributional assumptions and therefore it does not 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.

Let us demonstrate our methodology, which we will refer to as *distinctive-collexeme analysis*, by means of one of the best known alternating pairs, the so-called 'dative alternation', i.e. the alternating pair consisting of the ditransitive

construction and the prepositional dative with *to*, as exemplified in (1) above and repeated here as (6):

(6) a. *Ditransitive*: John sent Mary the book.
b. To-*dative*: John sent the book to Mary.

There is a number of verbs that occur in both of these constructions, which may lead us to assume that the two constructions are semantically equivalent (this is, of course, the reason why linguists think of them as an 'alternation' in the first place). However, there is also a number of differences between the two constructions in terms of the semantic restrictions they place on the verbs and NPs that can occur in them. An analysis of the verbs that are distinctive for each construction (i.e. that distinguish it significantly from the other) may help us elucidate the existence and degree of fine semantic differences between the two that might, in turn, explain the different restrictions.

In order to calculate the distinctiveness of a given collexeme, we need four frequencies: the lemma frequency of the collexeme in construction A, the lemma frequency of the collexeme in construction B, and the frequencies of construction A and construction B with words other than the collexeme in question. These can then be entered in a 2-by-2 table and submitted to the Fisher exact test (or any other distributional statistic). Obviously, defining what counts as an instance of construction A and construction B may involve decisions on the part of the researcher that have to be justified on theoretical grounds.

Table 1 shows the frequencies required for a distinctive collexeme analysis of the verb *give* in the ditransitive and the *to*-dative (for expository purposes, it also gives the expected frequencies for each combination of verb and construction in parentheses). The figures in italics were derived from a corpus (the ICE-GB, see further below), the other figures are the results of additions and subtractions.

The p-value resulting from the computation of Fisher exact for this distribution is exceptionally small: p=1.84E-120. This tells us that the verb *give* is

Table 1. The distribution of *give* in the ditransitive and the *to*-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

highly distinctive for one of the two constructions, but it does not tell us for which one. In order to determine this, we need to compare the observed frequencies with the expected frequencies. As the expected frequencies provided in parentheses show, *give* occurs more than twice as often in the ditransitive and less than half as often in the *to*-dative than would be expected given a random distribution. Thus, although *give* occurs in both constructions, it is highly distinctive for the ditransitive (in fact, as we will show below it is the *most* distinctive collexeme). Of course, this finding in isolation tells us relatively little about the semantics of the two constructions; the potential of the present methodology unfolds only when we repeat the analysis for all verbs occurring in the two constructions and rank them according to their distinctiveness value.[2]

3. Case studies

In the following sub-sections, we will apply distinctive-collexeme analysis as described in the previous section to five well-known and much investigated alternating pairs and show how the results of such an analysis bear on some of the issues discussed with respect to each of the alternating pairs in the literature, with a focus on the issues raised at the end of Section 1.

All studies are based on the British component of the *International Corpus of English* (ICE-GB), a corpus manually annotated for parts of speech, grammatical relations, syntactic (tree) structure and a variety of morphosyntactic features such as tense, aspect, mood, voice, transitivity etc. In each case, the formal configuration of morphosyntactic elements corresponding to the construction in question was extracted from the corpus with the help of a grep tool, making full use of the grammatical markup provided. In some cases, this was an entirely straightforward matter. For example, all ditransitive clauses in the ICE-GB are annotated as <*ditr*> and, thus, could be extracted in a single pass; the markup of the ICE-GB allowed a similarly straightforward extraction of verb-particle constructions, actives and passives, the *will*-future and the *going-to* future and the *s*-genitive and the *of*-genitive. Only in one case, heavy manual post-editing was necessary: in order to extract the *to*-dative, we searched for all clauses containing a direct object and a PP with *to* (for actives) and all clauses containing a verb annotated as <*pass*> and containing a PP with *to* for passives. Obviously, this search found all instances of the *to*-dative, but it also turned up a vast number of false positives which had to be manually discarded, for example, clauses with purpose or true locative adverbials. Both authors edited the

search results independently to ensure maximum accuracy.[3] Thus, for all of the constructions considered here, we can be sure to have identified all occurrences (barring annotation errors).

Once we had retrieved all instances of the alternating pair under investigation, the collexeme tokens (i.e. the words occurring in the slot under investigation) were lemmatized, and the frequency of each lemma was determined. The frequencies for each word were cross-tabulated with the frequencies of the constructions as shown in Table 1 above and submitted to the Fisher exact test (Fisher exact p-values were computed using the *dhyper* function in the current version of R). Finally, the words were ranked according to their Fisher-exact value for each of the two constructions depending on the direction in which they deviated from the expected frequency.

3.1 The dative alternation

Let us begin with what is probably the most widely discussed case of constituent-order alternation in English, the dative alternation already mentioned in the previous section (i.e. the ditransitive construction in (7a) and the *to*-dative construction in (7b); note that in this paper, we are not concerned with the prepositional dative with *for*):

(7) a. [$\text{SUBJ}_{\text{agent}}$ V $\text{OBJ}_{\text{recipient}}$ $\text{OBJ}_{\text{theme}}$]
e.g. *John sent Mary the book*.
b. [$\text{SUBJ}_{\text{agent}}$ V $\text{OBJ}_{\text{theme}}$ to-$\text{ADV}_{\text{recipient}}$]
e.g. *John sent the book to Mary.*

This alternating pair has spawned a vast body of research concerned with the formal, functional, and semantic issues discussed in Section 1.

First, a number of studies in the transformational-generative tradition have tried to determine which of the two constructions is basic and which is derived; in these studies, the two constructions are seen as purely syntactic alternatives that are semantically and pragmatically equivalent (see, for example, Fillmore (1965), who considers the prepositional dative the basic form from which the ditransitive is derived transformationally, or Burt (1971), who argues in favor of the opposite transformational direction).

Second, a number of functional or performance-oriented studies have argued that information structure is the primary factor determining the choice between the two constructions, claiming that speakers choose a member of the alternating pair based on whether its postverbal arguments, in a given context,

best adhere to linear-precedence principles related to topicality (e.g. Erteschik-Shir 1979: 443, 449ff.; Thompson 1990: 245f.; Goldberg 1995: 91ff.) or length (e.g. Hawkins 1994: 213).

Third, and most important for our purposes, a number of studies have focused on the semantics of the two constructions as the primary determining factor. For example, Thompson and Koide (1987: 400) claim that the ditransitive is used where the distance between agent and recipient is small, and the *to*-dative where this distance is large (in their view, this distance is iconically reflected in the distance between the subject of the construction and the argument encoding the recipient – the first object in the ditransitive, and the *to*-adverbial in the *to*-dative). Other researchers have posited distinct constructional semantics for the two constructions (e.g. Goldberg 1995: Section 2.3.1, Chapters 5–7; Pinker 1989), claiming that the ditransitive means something like 'X causes Y to receive Z' and the *to*-dative something like 'X causes Z to move to Y', specified by some authors as 'continuous causation of accompanied motion' (Gropen et al. 1989: 243f.). Note that the suggested constructional meanings are fully compatible with the differences in distance between agent and recipient postulated by Thompson and Koide (1987); 'causing Y to receive Z' does not suggest a long distance between agent and recipient, while 'causing Z to move to Y' does.

The predictions made by these approaches concerning distinctive collexemes are relatively straightforward: according to purely syntactic or purely information-structural approaches, one would predict that the two constructions should not differ at all with respect to their preferred verbs. The semantic approaches predict that such differences should exist, and that the ditransitive should prefer verbs of direct face-to-face transfer, while the *to*-dative should prefer verbs of transfer over a distance. Consider Table 2, which lists the results of the distinctive-collexeme analysis (to avoid various complications arising from the fact that there are verbs that do not alternate between the two constructions under any circumstances, we included only verbs that occur at least once in each construction in the ICE-GB; we also included only ditransitives with nominal objects, discarding those with sentential ones).

The results broadly support a semantic approach, in that there clearly are collexemes distinguishing between the ditransitive and the *to*-dative. Furthermore, the specific suggestions concerning the meaning of the two constructions are largely confirmed. For the ditransitive, we find that the most distinctive collexeme is *give*, a verb that matches the proposed constructional semantics of 'causing to receive' perfectly (note also that *give* is the strongest overall collex-

Table 2. Collexemes distinguishing between the ditransitive and the *to*-dative

Ditransitive (N=1,035)		*To*-dative (N=1,919)	
Collexeme	*Distinctiveness*	*Collexeme*	*Distinctiveness*
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)	0.0002
offer (43:15)	9.95E-10	pass (2:29)	0.0002
cost (20:1)	9.71E-09	make (3:23)	0.0068
teach (15:1)	1.49E-06	sell (1:14)	0.0139
wish (9:1)	0.0005	do (10:40)	0.0151
ask (12:4)	0.0013	supply (1:12)	0.0291
promise (7:1)	0.0036	read (1:10)	0.0599
deny (8:3)	0.0122	hand (5:21)	0.0636
award (7:3)	0.0260	feed (1:9)	0.0852
grant (5:2)	0.0556	leave (6:20)	0.1397
cause (8:9)	0.2131	keep (1:7)	0.1682
drop (3:2)	0.2356	pay (13:34)	0.1809
charge (4:4)	0.2942	assign (3:8)	0.4243
get (20:32)	0.3493	set (2:6)	0.4267
allocate (4:5)	0.3920	write (4:9)	0.4993
send (64:113)	0.4022	cut (2:5)	0.5314
owe (6:9)	0.4369	lend (7:13)	0.5999
lose (2:3)	0.5724		

eme for this construction, cf. Stefanowitsch & Gries (2003)). *Give* and most other distinctive collexemes encode transactions that involve a direct contact between agent and recipient; these transactions may be literal (as in the case of *give*) or they may be metaphorical, instantiating one of the semantic extensions postulated by Goldberg (1995: Section 2.3.1). For example,

– *tell*, *teach* and *ask* instantiate the COMMUNICATION AS TRANSFER metaphor – note that in their typical senses these verbs encode interpersonal communication without an intervening medium;
– *show* instantiates the PERCEIVING AS RECEIVING metaphor – note that it involves visibility between show-er and show-ee;
– *offer*, *promise* instantiate the SATISFACTION CONDITION extension (i.e. the satisfaction conditions of the speech acts referred to by these verbs imply the basic literal meaning);
– *deny* instantiates the CAUSE NOT TO RECEIVE extension (i.e. it is the negation of the basic literal meaning).

For the *to*-dative, we find that *bring* is the most distinctive collexeme; again, this verb matches the proposed constructional meaning of '(continuously) caused (accompanied) motion' perfectly. *Bring* and other verbs strongly distinctive for the *to*-dative, such as *take* or *pass*, also involve some distance between agent and recipient that must be overcome to complete the action denoted by the verb. Second-ranked *play* also belongs into this category; its distinctiveness may seem surprising out of context, but it is due to the large number of uses in the context of sports commentary, such as *Michalichenko plays [the ball] forward to the halfway line* (ICE-GB S2A-014 #145:1), and is thus at least in part due to the characteristics of the corpus we used.

The suggested constructional meanings do not straightforwardly account for all differences between the two constructions, however. For example, it is interesting to note that most commercial transaction verbs, such as *sell*, *supply* and *pay* are all distinctive for the *to*-dative rather than the ditransitive – the only exception to this generalization is *cost*. This finding would have been difficult to anticipate since a commercial transaction frame seems to be semantically more compatible with the ditransitive; it involves a change of possession, but not necessarily a change of location (consider, for example, the context of selling a house). On the other hand, one might argue that these verbs *typically* involve a movement of the commodity to the buyer or the paid sum to the seller (contrast *cost*, which never involves motion, and thus would be expected to occur in the ditransitive). Such observations may offer a useful starting point for a more refined analysis of the two constructions' semantics.

In the context of studying alternating pairs of constructions, it may of course also be useful to look at verbs that are *not* distinctive for either construction, i.e. verbs that not only *can* occur readily in both constructions, but actually *do*. In the case of the ditransitive and the *to*-dative, the verbs which alternate most freely are *lend* (7:13), *send* (64:113), *get* (20:32) and *write* (4:9). Especially for *send* and *write*, it is easy to see why these are prime candidates for a relatively free alternation between the two constructions: when used in either of the two constructions, their meanings involve both the 'transfer' meaning of the ditransitive and the 'caused motion' meaning associated with the *to*-dative. In the case of such freely alternating verbs, the specific construal of the events in question will determine the choice of construction (for example along the lines suggested by Thompson & Koide 1987).

3.2 Active and passive

The most famous alternating pair in English is probably the one consisting of active and passive voice, as in (8):

(8) a. [SUBJ V OBJ]
e.g. *Picasso painted this picture.*
b. [SUBJ *be* V-*ed* (*by*-ADV)]
e.g. *This picture was painted by Picasso.*

Again, the relationship between active and passive has been discussed as a purely syntactic operation (e.g. in early transformational grammar, cf. Chomsky 1957), or as a way of adjusting the information structure of an utterance (cf. e.g. Givón 1993:47, who claims that the active voice encodes the typical case where the Agent is the most topical referent, while the passive voice is used to place the Patient in the topic position in those cases where the Patient is the most topical referent). There is no doubt that the adjustment of information structure is one important function of the passive (cf. also Biber et al. 1999:941; Huddleston & Pullum 2002:1444), but as the passive does not apply freely to all syntactically transitive verbs, additional restrictions need to be recognized. Several studies have shown that these restrictions are largely semantic in nature (cf. e.g. Bolinger 1975; Rice 1987), but few of these studies have explicitly characterized the meaning of the passive construction independently of that of the active transitive construction. One such characterization is given in Pinker (1989), who defines its meaning as follows (where *X* is the referent of the passive subject and *Y* is the referent of the implicit second core argument, which may be expressed in a *by*-phrase):

> X is in the circumstance characterized by Y's acting on it (more generally, the circumstance for which Y is responsible [...]). (Pinker 1989:91)

Again, the predictions made by the different views of the active-passive distinction with respect to distinctive collexemes are relatively straightforward: if it is a purely syntactic alternation, or if its function is exclusively one of adjusting information flow, we would not expect there to be any strong collexemes (since the verbs in question would occur equally naturally in the active voice and in the passive voice). If semantics plays a role in the choice of construction, we would expect there to be semantically motivated classes of distinctive collexemes. More specifically, if Pinker's characterization is correct, the active-voice construction should prefer verbs whose active-voice direct objects are not (eas-

Table 3. Collexemes distinguishing between active and passive

Active transitive (N=53,144)		Passive (N=11,912))	
Collexeme	*Distinctiveness*	*Collexeme*	*Distinctiveness*
have (3957:1)	0	base (9:125)	1.97E-80
think (2319:19)	2.38E-175	concern (28:100)	6.92E-49
get (1929:5)	2.88E-162	use (786:377)	1.74E-31
say (1916:58)	1.99E-101	involve (117:122)	2.38E-30
want (1106:1)	1.76E-96	publish (35:68)	2.37E-26
do (2391:182)	1.20E-62	associate (16:53)	6.21E-26
know (1344:53)	4.53E-62	bear (46:70)	1.19E-23
see (1407:72)	1.71E-54	engage (8:41)	5.23E-23
mean (611:7)	5.78E-44	design (30:56)	1.60E-21
like (503:1)	4.00E-43	confine (1:27)	2.79E-19
try (584:8)	1.45E-40	entitle (1:27)	2.79E-19
hope (269:6)	5.06E-17	relate (17:42)	1.00E-18
believe (294:9)	1.89E-16	deposit (1:22)	1.12E-15
remember (271:8)	1.71E-15	compare (34:45)	1.98E-14
feel (256:11)	2.90E-12	derive (10:29)	3.68E-14
suppose (138:1)	1.93E-11	deal (1:18)	8.32E-13
wish (132:1)	6.23E-11	aim (18:32)	1.34E-12
thank (114:0)	9.49E-11	release (25:35)	5.94E-12
enjoy (135:2)	4.51E-10	attach (20:32)	6.33E-12
ensure (101:0)	1.32E-09	store (11:26)	6.75E-12

ily) construable as patients, i.e. stative verbs; in contrast, the passive-voice construction should prefer verbs encoding actions with a salient and relatively permanent end-state. Table 3 lists the results of the distinctive-collexeme analysis (based on all actives and passives of transitive verbs in the ICE-GB).

The results show that constructional semantics is an influencing factor in the choice between active and passive voice: again, there are clearly distinctive collexemes for each. These also support the more specific suggestions discussed above. With respect to active voice, we find that *have* is the most distinctive collexeme; a paradigm case of a stative verb (and a frequently cited example for a transitive verb that never passivizes in its literal use). The other distinctive collexemes are also stative verbs; all of them referring to mental or emotional states. There is a number of verbs that do not fit this characterization; we will presently return to these.

The distinctive collexemes for the passive construction clearly confirm Pinker's characterization, and thus the claim that it is a primarily semantic construction, on a par with argument-structure constructions like the ditran-

sitive or the *to*-dative. The verbs overwhelmingly encode processes that cause the patient to come to be in a relatively permanent end state; these end states are often more naturally encoded by stative passives (as in *The results [are] based on a limited sample of patients* (S2A 033: #4:1:A)) than by regular passives (as in *The results were based on a limited number of patients (by the researchers)*). Actually, for some of the verbs in question the end state is so much more salient than the process that led to it that it is not clear that their passive uses are still capable of being perceived as passives at all. For example, passive uses of *base*, *concern*, and *involve* may well be felt to be copular constructions containing deverbal adjectives (cf. in this context Quirk et al.'s (1985: Section 3.74–78) discussion of their notion of 'passive gradient').

Let us briefly return to those verbal collexemes that are distinctive for the active voice even though they do not fit the majority pattern of stative verbs (*get*, *say*, *do*, *mean*, *try*, *thank*, *ensure*). These are like the typical passive verbs in that they encode actions rather than states, but they are different in that the actions they encode do not lead to permanent end states. The reason that they are distinctive for the active voice is thus presumably their incompatibility with the meaning of the passive.[4]

In sum, the distinctive-collexeme analysis shows that passive voice is a construction in its own right with its own specific semantics. It encodes a situation where the referent of the passive-voice subject has come to be in some relatively stable end state as a result of someone acting on it. The distinctive collexemes of the active-voice construction are either action verbs that do not lead to such end states, or they are states that are not brought about by someone acting on the entity in this state. The highly dynamic action verbs thought of as typical for active and passive sentences occur in both constructions equally frequently, and may thus be typical for both constructions, but are not distinctive for either of them.

3.3 Verb-particle constructions

Let us now turn to an alternating pair of constructions that, in contrast to the previous two cases, differ exclusively in the order of constituents – the transitive verb-particle constructions exemplified in (9):

(9) a. [SUBJ V Particle OBJ]
e.g. *John picked up the book.*

b. [SUBJ V OBJ Particle]
e.g. *John picked the book up.*

As with the dative alternation, a number of studies have dealt either with the issue which of these constructions is basic and which is derived (cf. Legum 1968), or with the factors that motivate speakers to choose one or the other. With respect to the latter issue, the factors that have been suggested include those also proposed for the dative alternation and/or the active-passive pair, i.e. the degree of topicality and the length of the direct object; in addition, it has been suggested that the degree of idiomaticity of the transitive phrasal verb has an influence (cf. Gries 2003a for comprehensive discussion of all theses variables and their influence). With respect to these three variables, it has been claimed repeatedly that the order [V Prt OBJ] is preferred with non-topical and/or long object NPs and with idiomatic verb-particle combinations, while the order [V OBJ Prt] is preferred with topical and/or short object NPs and with non-idiomatic (literal or metaphorical) verb-particle combinations, where the particle denotes the spatial goal or the resultant state of the direct object's referent.

While factors like topicality or length of the object would not show up in a distinctive-collexeme analysis, idiomaticity should, as should other semantic factors that might lead to preferences for individual verb-particle combinations. Very little is known about whether such preferences exist, but Browman (1986) suggests they may, showing that *up* in general and the transitive phrasal verb *pick up* in particular prefer the order [V Prt OBJ] (Browman 1986: 317). Thus, the predictions are as before: if we are dealing with a purely syntactic or information-structural difference, there should be no distinctive collexemes; if semantic factors (such as idiomaticity) play a role, such collexemes should exist. Table 4 lists the results of the distinctive-collexeme analysis.

The results clearly show an influence of semantics. Beginning with [V Prt OBJ], even a superficial look at its distinctive verbal collexemes underscores the relevance of idiomaticity. For example, none of the uses of first-ranked *carry out* in the ICE-GB has a literal spatial meaning (as in *John carried his dog out*); they can all be paraphrased by *perform* or *execute*. The same is true of most other verbs on the list – the particle does not literally denote the endpoint of a path for any of them. In a few cases, however, the verb-particle combinations instantiate metaphorical mappings from the source domain of literal spatial motion to a more abstract target domain; examples include *A bad time to bring out a war film* (S2B-033 #43:1:A) and *The use of large quantities of straw [...]*

Table 4. Distinctive collexemes for *[V Prt Obj]* and *[V Obj Prt]*

V Prt Obj (N=1,251)		V Obj Prt (N=1,192)	
Collexeme	*Distinctiveness*	*Collexeme*	*Distinctiveness*
carry out (49:1)	9.10E-14	get back (0:18)	2.30E-06
find out (49:5)	3.83E-10	get out (2:21)	1.91E-05
point out (43:3)	4.42E-10	play back (1:12)	0.0013
set up (42:8)	1.06E-06	turn off (2:14)	0.0015
take on (37:7)	4.60E-06	ring up (3:16)	0.0015
build up (18:1)	5.44E-05	get on (0:7)	0.0065
take up (35:9)	8.76E-05	get together (0:7)	0.0065
give up (18:3)	0.0010	get in (4:15)	0.0070
work out (20:4)	0.0011	let down (0:6)	0.0134
set out (10:0)	0.0012	get down (0:5)	0.0275
bring about (10:1)	0.0072	have back (0:5)	0.0275
bring out (12:2)	0.0081	have on (0:5)	0.0275
make out (7:0)	0.0092	play forward (0:5)	0.0275
wipe out (6:0)	0.0179	play out (0:5)	0.0275
play down (6:0)	0.0179	trace back (0:5)	0.0275
cut down (6:0)	0.0179	turn round (0:5)	0.0275
fill in (13:4)	0.0304	phone up (1:7)	0.0300
top up (5:0)	0.0351	send back (1:7)	0.0300
lay down (9:2)	0.0387	take off (4:12)	0.0306
rule out (13:5)	0.0586	take out (15:26)	0.0413

also cuts down the smell (W2B-027 #72:1). In the case of *find out* and *point out*, there is an additional syntactic factor that strengthens their association to [V Prt OBJ]: they often take *that*-clauses as objects, which can never occur with the alternative order.

A look at the distinctive collocates of the alternative order [V OBJ Prt] also confirms the influence of semantics; as expected, it occurs predominantly with non-idiomatic verb-particle combinations where the particle denotes a spatial goal or a result. For example, nearly all instances of first-ranked *get back* in the ICE-GB encode situations where the (concrete) referent of the direct object NP moves to the spatial location referred to by the particle (as in *Why did he get the money back* (S1B-005 #61:1:A)); the same holds for second-ranked *get out* (cf. *[When does the] library shut because I want to get a book out overnight* (S1A-069 #224:2:B)) and third-ranked *play back* (cf. *It's with Vasili Khulkov who plays the ball back to Galiamin* (S2A-010 #2:1:A)). In other cases, such as fourth-ranked *turn off*, the particle encodes a resultant state that could be referred to by the

same particle in other constructions (for example, if you *turn something off*, then it *is off*).[5]

As before, the predictions from the literature do not exhaust an analysis based on distinctive collexemes. For example, *ring up* and *phone up* constitute an exception to the general pattern in that they are idiomatic, i.e. their particle does not describe a spatial goal or resultant state (instead, some authors have claimed, it conveys a kind of perfective meaning). Given that they are synonyms, it seems unlikely that their unexpected behavior is an accident; it is likely that a more detailed investigation of a larger corpus would allow us to determine whether they are part of a systematic semantically based exception (note, for example, that they are the only verbs of communication among the distinctive collexemes for either construction).

Finally, let us briefly point out that the technique of distinctive collexemes can of course also be applied to the particle alone; we then find that *out* and *up* (which, as mentioned, typically receive a relatively abstract, perfective interpretation in the verb-particle constructions) are significantly distinctive for [V Prt NP] whereas *back*, *round*, *together*, *forward* and *through* (which retain their spatial meaning in the verb-particle constructions) are significantly distinctive for [V NP Prt].

3.4 Will *vs.* be going to

As mentioned in the Introduction, the method of distinctive collexemes is not restricted to the classic cases of 'alternations'; it can be applied to any pair of constructions expressing roughly the same meaning, for example, the two major future tense constructions in English:

(10) a. [SUBJ *will* VP]
e.g. *John will send Mary a book.*
b. [SUBJ *be going to* VP]
e.g. *John is going to send Mary a book.*

Unlike the previously discussed cases, the choice between these constructions has not been a major issue in linguistic theorizing, but it is a recurrent topic in English student's grammars, which tend to focus on three differences between the two: (i) when talking about one's own future actions, *be going to* is used for more planned, 'premeditated' actions than *will* (e.g. Thompson & Martinet 1986: 185; Murphy 1986: 16); (ii) when talking about future events not involving oneself, *be going to* expresses a greater certainty on the part of the speaker

Table 5. Collexemes distinguishing between the *will* and the *be going to* future

WILL (N=3,667)		BE GOING TO (N=980)	
Collexeme	*Distinctiveness*	*Collexeme*	*Distinctiveness*
see (90:8)	0.0004	say (28:42)	1.12E-12
find (58:4)	0.0015	do (105:68)	2.02E-08
give (69:7)	0.0047	happen (11:15)	4.77E-05
know (34:2)	0.0108	have (126:61)	0.0001
provide (17:0)	0.0177	go (96:48)	0.0004
depend (15:0)	0.0285	win (1:6)	0.0005
want (22:1)	0.0305	stay (8:10)	0.0014
receive (14:0)	0.0361	use (19:14)	0.0045
consider (13:0)	0.0458	buy (2:5)	0.0059
remain (13:0)	0.0458	talk (16:11)	0.0160
become (19:1)	0.0553	show (22:13)	0.0213
finish (12:0)	0.0581	get (84:34)	0.0275
hold (11:0)	0.0736	suggest (1:3)	0.0315
include (11:0)	0.0736	be (664:203)	0.0357
notice (11:0)	0.0736	put (19:11)	0.0362
follow (10:0)	0.0934	invest (0:2)	0.0444
reach (10:0)	0.0934	measure (0:2)	0.0444
need (16:1)	0.0985	perform (0:2)	0.0444
send (21:2)	0.1080	photocopy (0:2)	0.0444
accept (9:0)	0.1184	rehearse (0:2)	0.0444

than *will* (Thompson & Martinet 1986: 185ff.; Murphy 1986: 16), and, perhaps related to this point, (iii) *be going to* is used for talking about a more immediate future than *will* (Thompson & Martinet 1986: 185; Murphy 1986: 16). Clearly, none of these factors can straightforwardly be related to particular (semantic classes of) verbs, and thus they cannot be expected to show up transparently in the lists of distinctive collexemes. However, some reflex of these factors might be expected. Table 5 lists the results of the distinctive-collexeme analysis.

The most striking difference between the two lists concerns the dynamicity of the actions and events encoded. The distinctive collexemes for *will* are overwhelmingly relatively non-agentive or low-dynamicity actions (*find*, *receive*, *hold*, *finish*, *reach*) including perception/cognition events (*see*, *know*, *want*, *consider*, *notice*, *need*, *accept*), or states (*depend*, *remain*, *become*). Only five of the top 20 collexemes encode dynamic actions (*give*, *provide*, *include*, *follow*, *send*). With *be going to*, the situation is reversed: only five of the top 20 collexemes encode states or non-agentive actions (*have*, *stay*, *be*, *happen*, *get*); the other fifteen collexemes encode very dynamic actions. A second difference

concerns the specificity of the actions and events involved: the list for *be going to* contains some very specific actions (*invest*, *measure*, *photocopy*), and seems to encode more specific actions and events than *will* in general.

In other words, *be going to* encodes more dynamic and more specific actions and events than *will*. The higher specificity (to the extent that it can be argued to be present) may be related to the greater immediacy and certainty claimed to be associated with *be going to*. The higher dynamicity may be related to the notion of premeditation; more dynamic actions require more effort, and hence perhaps more planning. However, the results of the distinctive collexeme analysis are interesting even if we do not relate them to previous analyses, but simply add 'degree of dynamicity' as an additional distinguishing factor.

3.5 *s*-genitive vs. *of*-construction

Finally, let us turn to a famous alternating pair of constructions at the NP level, the case of the *s*-genitive and the so-called *of*-genitive (or *of*-construction), shown in (11):

(11) a. [$NP_{modifier}$'*s* N_{head}]
e.g. *the university's budget*
b. [DET N_{head} of $NP_{modifier}$]
e.g. *the budget of the university*

It has been widely assumed that these two constructions are semantically equivalent and that their distribution is governed by linear precedence preferences related to givenness (e.g. Standwell 1982; Osselton 1988; cf. Quirk et al. 1985: §17.45), animacy (e.g. Jespersen 1949; Hawkins 1981; cf. also Quirk et al. 1985: §17.39), or a combination of the two (Deane 1987). The basic idea behind such accounts is that NPs with given/animate referents precede those with new/inanimate referents, and that therefore the *s*-genitive is chosen when the modifier is more given/animate, while the *of*-construction is chosen when the head is more given/animate. Determining the givenness status of a noun requires an assessment of its context, and thus givenness is not amenable to distinctive-collexeme analysis; in contrast, animacy can be read off a wordlist, and thus is a perfect candidate for this method. As previous studies have shown that givenness is not a factor anyway (Altenberg 1980; cf. Gries 2002; Stefanowitsch 1998, 2003), we will focus on animacy here.

The predictions of linear-precedence accounts of *s*-genitive and *of*-construction concerning the two constructions' distinctive collexemes are straightfor-

Table 6. Modifiers in the *s*-genitive and the *of*-construction

Genitive mod (N=9,892)		Of-construction mod (N=11,103)	
Collexeme	*Distinctiveness*	*Collexeme*	*Distinctiveness*
he (1664:0)	0	thing (0:101)	9.17E-29
they (1415:0)	0	war (0:88)	3.82E-25
you (1291:6)	0	life (0:75)	1.58E-21
I (1221:0)	0	time (1:68)	4.62E-18
we (697:0)	2.31E-234	system (2:71)	1.24E-17
she (683:0)	1.50E-229	state (0:50)	1.39E-14
it (790:113)	1.98E-149	work (0:47)	9.47E-14
pers. names (781:155)	8.25E-123	force (0:41)	4.38E-12
today (13:0)	5.62E-05	society (1:43)	2.60E-11
mother (19:3)	0.0002	century (0:37)	5.63E-11
BBC (11:0)	0.0003	house (0:34)	3.83E-10
women (11:0)	0.0003	<amount>pound (0:33)	7.24E-10
widow (10:0)	0.0005	people (38:122)	8.11E-10
Iraq (22:6)	0.0007	area (0:32)	1.37E-09
Britain (31:14)	0.0026	this (0:32)	1.37E-09
yesterday (11:2)	0.0064	data (0:31)	2.60E-09
Lord (14:4)	0.0081	information (0:31)	2.60E-09
tomorrow (6:0)	0.0109	interest (0:31)	2.60E-09
mum (5:0)	0.0232	that (0:31)	2.60E-09
father (11:4)	0.037	water (0:31)	2.60E-09
boy (4:0)	0.0493	population (0:30)	4.92E-09
IBM (4:0)	0.0493	material (0:27)	3.34E-08
Observer (4:0)	0.0493	scheme (0:27)	3.34E-08
professor (4:0)	0.0493	law (0:26)	6.32E-08
taxpayer (4:0)	0.0493	order (0:25)	1.2E-07

ward: the *s*-genitive should attract animate modifiers and inanimate heads, and the associations should be reversed for the *of*-construction. Table 6 shows the distinctive collexemes for the modifier position of the two constructions.

Clearly, the prediction is borne out spectacularly. Twenty-two of the top 25 distinctive collexemes for the *s*-genitive refer to human beings; the pronouns come out at the top, followed by personal names (which we grouped together as a single lemma), followed by kinship terms and metonymic references via organizations (*BBC*, *IBM*) or countries (*Iraq*, *Britain*) and other human nouns. The three exceptions are the temporal nouns *today*, *yesterday*, and *tomorrow* (we will return to these immediately). The top 25 distinctive collexemes of the *of*-construction, in contrast, are mainly made up of abstract nouns and a few concrete inanimate nouns. There are some directly or metonymically hu-

Table 7. Heads in the *s*-genitive and the *of*-construction

Genitive head (N=10,440)		Of-construction head (N=11,982)	
Collexeme	*Distinctiveness*	*Collexeme*	*Distinctiveness*
friend (125:8)	3.08E-32	sort (7:253)	3.12E-59
mother (92:0)	2.32E-31	part (22:252)	8.46E-45
father (79:1)	2.20E-25	kind (10:206)	1.04E-43
life (129:22)	2.05E-23	number (27:218)	2.01E-33
mind (77:5)	3.02E-20	end (8:155)	1.06E-32
letter (80:11)	1.33E-16	amount (1:100)	2.40E-26
wife (50:1)	6.57E-16	type (1:90)	1.18E-23
hand (66:9)	5.69E-14	rest (0:57)	2.89E-16
correspondent (43:1)	1.22E-13	matter (0:44)	1.02E-12
husband (38:0)	2.34E-13	member (18:98)	1.42E-12
parent (45:2)	3.52E-13	edge (1:45)	1.20E-11
mum (37:0)	5.04E-13	form (18:92)	2.24E-11
work (76:17)	2.69E-12	piece (4:55)	2.38E-11
colleague (35:1)	4.60E-11	level (7:63)	4.38E-11
eye (51:7)	4.71E-11	example (1:42)	7.40E-11
job (41:3)	5.08E-11	side (21:96)	7.49E-11
foot (36:2)	2.29E-10	range (5:54)	2.35E-10
sister (29:0)	2.31E-10	area (11:70)	2.74E-10
child (37:3)	8.18E-10	middle (0:35)	2.92E-10
speech (31:1)	8.78E-10	group (7:59)	3.55E-10
party (30:1)	1.83E-09	aspect (1:39)	4.54E-10
way (64:17)	2.08E-09	loss (1:38)	8.30E-10
book (56:13)	3.39E-09	series (2:41)	1.40E-09
place (38:5)	1.11E-08	copy (1:36)	2.77E-09
name (79:29)	1.72E-08	nature (5:49)	3.48E-09

man nouns, but these refer to relatively abstract or unspecific categories (*state*, *society*, *people*, *population*).

However, with respect to the head nouns, the predictions of the linear-precedence accounts are not borne out at all. Table 7 lists the distinctive collexemes.

The distinctive collexemes for the *s*-genitive are again predominantly animate nouns. Most of these refer to humans in relation to other humans, either in the context of kinship (*mother*, *father*, *wife*, *husband*, etc.), or in the context of social or workplace relations (*friend*, *correspondent*, *colleague*). Some inanimate nouns occur on this list; these are body parts (*hand*, *eye*, *foot*, and perhaps *mind*), nouns referring to events/actions and their results (*life*, *letter*, *work*, *speech*), and possessions (*book*, *place*, and perhaps *name*). The distinc-

tive collexemes of the *of*-construction are, again, mainly abstract or inanimate nouns referring to taxonomic relations (*sort*, *kind*, *type*, *example*) or parts and quantities (*part*, *number*, *end*, *amount*, *rest*, etc.). The two animate nouns that appear here all fall into the latter group as well (*member*, *group*). These results clearly argue against a linear precedence approach. Instead, they support semantic accounts, such as those developed in Langacker (1995) and Stefanowitsch (1998, 2003), which claim that the *s*-genitive and the *of*-construction are actually two semantically distinct constructions. Stefanowitsch (2003) argues that the *s*-genitive essentially encodes kinship/social relations and possession, and the *of*-construction encodes taxonomic and meronymic relations. This also accounts for the seemingly correct predictions of the linear-precedence accounts with respect to the modifiers: animate nouns are simply more likely to appear in possessive/kinship relations, while abstract and inanimate nouns are more likely to be quantified and taxonomized. A constructional account can also accommodate the exceptional temporal nouns in the head slot of the *s*-genitive, which can be argued to be instances of a particular sub-construction of the *s*-genitive (cf. Stefanowitsch 2003: 440, Note 8).

In sum, the distinctive-collexeme analysis of the *s*-genitive and the *of*-construction yields solid evidence for previously suggested semantics-based accounts of these two constructions; they are in a paraphrase relationship only under very specific circumstances (cf. Stefanowitsch 2003), but the analysis presented here shows that such cases play a marginal role in the characterization of the constructions.

4. Methodological issues

As the case studies in Section 3 have shown, distinctive-collexeme analysis yields a substantial number of significantly distinctive collexemes for a range of alternating pairs, and these distinctive collexemes have a potential to reveal non-trivial properties of the constructions constituting these pairs. In this section, we will look at two fundamental properties of the method that we have so far ignored. First, we will address the question whether the method can serve as a basis for making above-chance-level predictions about which member of an alternating pair will be chosen for a particular verb. Second, we will address the possibility that distinctive collexeme analysis is overly sensitive, i.e. that it produces significantly distinctive collexemes even for alternating pairs where this would not be plausibly expected.

4.1 The predictive power of distinctive collexeme analysis

First, let us look at the degree to which significantly distinctive collexemes predict which member of an alternating pair native speakers will choose in a situation where both would be possible. For example, a speaker who is about to produce an utterance describing the successful manual transfer of an item X from a person A to another person B might want to use the verb *give* and must now choose whether to use the ditransitive or the *to*-dative; does the fact that he or she has chosen the verb *give* suffice as a basis for guessing with above-chance success which construction he or she will choose?

While many of the previous investigations of this alternating pair have yielded surprisingly good prediction accuracies (exceeding 80%, cf. Gries 2003a, b), we know of no previous work which took the notion of lexical bias serious enough to investigate to what degree the constructional choice can be predicted just on the basis of the verb. In what follows, we will remedy this oversight for each of the construction pairs discussed in Section 3. The first pair, the dative alternation, will be discussed in some detail to demonstrate the procedure.

As mentioned in Section 3.1 above, our study of the dative alternation includes only those verbs that occur at least once in each construction; the resulting data base consists of 40 verbs constituting 1,772 tokens of both constructions. Of these tokens, 1,248 (21 types) have a bias towards the ditransitive (for 951 tokens / 11 types, this bias is significant, for 297 tokens / 10 types, it is not significant). The remaining 524 tokens (19 types) are biased towards the *to*-dative (for 337 tokens / 8 types, this bias is significant, for 187 tokens / 11 types, it is not significant).

Thus, 79.8% of all tokens of significantly distinctive collexemes for the ditransitive did in fact occur in the ditransitive. Likewise, 89% of all tokens of significantly distinctive collexemes for the *to*-dative did occur in the *to*-dative. That is, if we know that the speaker will use one of the distinctive collexemes, 82.2% of the following constructional choices can be predicted correctly without including any other (syntactic or pragmatic) factor supposed to govern the choice between the two constructions; these findings are summarized in Figure 1.

The accuracy of chance-level predictions (i.e. predictions based on the overall construction frequencies in the corpus) would have been 65% (since the *to*-dative accounts for 65% of all tokens of the alternating pairs; cf. below). Thus, knowing the verb alone leads to a proportional reduction of error

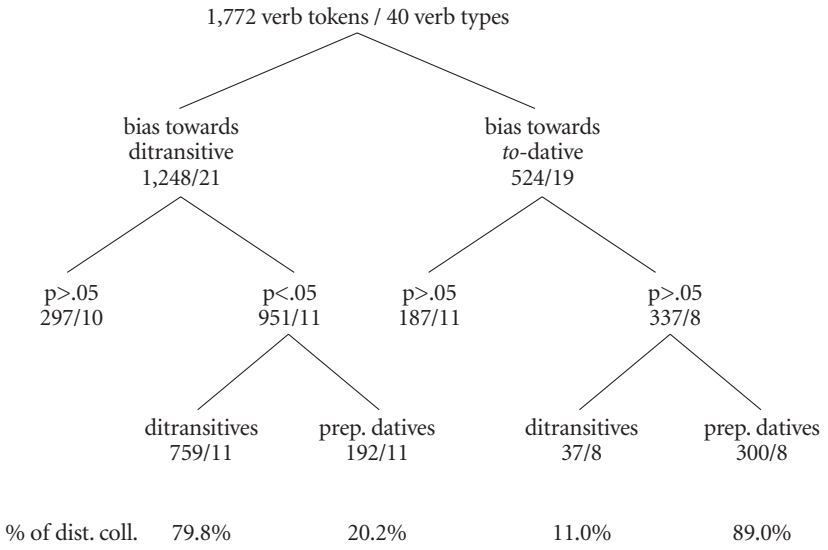

Figure 1. Distribution of verbs in the dative alternation

(PRE) of more than 17%. Table 8 shows the corresponding results for the other alternating pairs discussed in Section 3, calculated in the same way.

Interestingly, the results are not unequivocal: for three of the constructions, the distinctive collexemes allow us to predict the choice of construction above chance – for the active-passive pair and the two future tense constructions, however, predictions on the basis of distinctive collexemes are actually inferior to guesses on the basis of the overall construction frequencies with the significant collexemes. This raises two questions. First, is the technique overly sensitive in that it (sometimes) picks out words as significantly distinctive for a given construction that are not associated with that construction strongly enough to

Table 8. Prediction accuracies / PRE scores attained by distinctive collexeme analysis

alternating pair	% of distinctive collexemes (precision)		accuracy (combined)	% correct prediction by chance; PRE	
ditransitive – prep. dative	79.8	89	82.2	65	17.3
active – passive	95.9	57.9	76.9	81.7	–4.8
V Prt NP – V NP Prt	88.8	81.4	85.9	51.2	34.6
will – be going to	68.3	6	42.5	78.9	–36.5
's – of: head	86.5	89.2	87.9	53.4	34.5
's – of: modifier	96.6	95.9	96.3	52.9	43.4

make predictions about their distribution? Second, if this is the case, does this undermine the validity of the method?

We believe that both questions can be answered in the negative. With respect to the first question, consider where the accuracy of chance-level predictions comes from in the first place. It results from a strategy where we simply choose the more frequent member of a pair of alternating constructions, and predict for every token of every verb that it will appear in that construction. For example, there are 1,035 ditransitives and 1,919 *to*-datives in the ICE-GB, i.e. the *to*-dative accounts for 65% of all tokens of the pair. Our chance-level strategy would be to predict for every verb token that it will occur in the *to*-dative; these predictions will be accurate in 65% of all cases. In other words, the accuracy of chance-level predictions depends on the difference in frequency between the two members of the pair. If the two members of a pair are roughly equally frequent (i.e. entropy is high), chance-level predictions will be relatively inaccurate (this is the case for the dative alternation, the verb-particle constructions, and the genitives), and thus leave a wide margin for improvement. If the two members differ greatly in their frequency (i.e. entropy is low), chance-level predictions will be relatively accurate (this is the case for the two futures and the active-passive alternation), and thus there is a low margin for improvement. In purely quantitative terms, then, predictions based on distinctive collexemes (or any other non-chance criterion) are more likely to increase prediction accuracy for the first group of constructions than for the second, and this is, of course, exactly what we find.

However, in qualitative terms, the chance-level predictions are less telling even where they are more accurate than predictions based on distinctive collexemes; note that chance-level predictions are by their very nature extremely lopsided: they predict correctly for a large number of verb tokens that these will occur in the more frequent construction. However, they do not predict correctly for *any* verb token that it will occur in the less frequent construction. In other words, they simply ignore the less frequent member of the alternating pair completely, and thus make predictions that, quantitatively accurate though they may be, do not shed *any* light on the alternating pair.[6]

With respect to the second question, note that even if the discriminative potential of some verbs is not high enough to yield a uniformly high predictive power (for reasons just explained), this does not invalidate the semantic patterns revealed by the analysis, which sometimes tied in nicely with previous works and sometimes even unearthed hitherto unnoticed generalizations (cf. Stefanowitsch & Gries (2003) for more examples of such surprising findings).[7]

Thus, abandoning the technique just because it cannot predict constructional choices in all cases would be to throw out the baby with the bathwater by prematurely dispensing with a technique whose other merits are quite obvious.

4.2 Validation: *Try to* vs. *try and*

Potentially, distinctive-collexeme analysis may always yield distinctive collexemes, even for pairs of constructions where this is not plausibly expected. If this were the case, it would of course diminish the usefulness of the method to a certain degree. We therefore need to apply it to an alternating pair where we would not plausibly expect any distinctive collexemes, either because their meanings are too close, or because they are of a kind that is not easily reflected in classes of verbs.

One such pair is the one between the two main complementation patterns of *try*, [*try to* V] and [*try and* V].

(12) a. Let's try and keep it simple.
b. Let's try to keep it simple.

The choice between these two alternatives is often claimed to be mainly stylistically motivated (cf. e.g. the *Oxford English Dictionary*, s.v. *try*). Even where semantic differences have been proposed, they are very tenuous and of a kind that we would not expect to be reflected in particular verbs. For example, Nordquist (1998) argues that [*try and* V] indicates the speaker's belief that the agent is unlikely to be successful in achieving the event encoded in the complement of try, while [*try to V*] does not indicate any belief about the likely success of the agent. As 'likelihood of success' is not a notion encoded in individual verbs, we would not expect this meaning difference to show up in a collostructional analysis method). Table 9 shows the results of the distinctive-collexeme analysis.

As expected, there are almost no distinctive collexemes for this alternating pair. More precisely, there is only one significantly distinctive collexeme for each construction: *make* for [*try to* V] and *get* for [*try and* V]. Furthermore, even for these, the distinctiveness is weaker by many orders of magnitude than for the top collexemes of all other alternating pairs discussed here. Finally, the two distinctive collexemes are not interpretable in the context of the claims concerning stylistic or semantic differences between the two constructions.

Table 9. Collexemes distinguishing between *try to* and *try and*

Try to (N=103)		Try and (N=92)	
Collexeme	*Distinctiveness*	*Collexeme*	*Distinctiveness*
make (9:1)	0.0147	get (7:15)	0.0305
do (5:1)	0.1342	come (0:3)	0.1032
put (3:0)	0.1453	teach (0:2)	0.2213
analyse (2:0)	0.2777	show (0:2)	0.2213
give (2:0)	0.2777	learn (0:2)	0.2213
improve (2:0)	0.2777	set (1:2)	0.4575
persuade (2:0)	0.2777	see (1:2)	0.4575
use (2:0)	0.2777	bring (1:2)	0.4575
keep (3:1)	0.3542	be (1:2)	0.4575
absorb (1:0) + next 46 verbs	0.5282	win (0:1) + next 44 verbs	0.4718

5. Conclusion

This paper has introduced an extension of our previously proposed method of collostructional analysis specifically geared to investigating distinctive collexemes for pairs of constructions. This method can be applied in the context of what is traditionally thought of as 'alternation' (e.g. the dative 'alternation', particle 'movement', or the 'transformational' relationship between active and passive), but it can more generally be applied to any area of grammar where there is a choice between two more or less equivalent constructions (as in the case of *will*-future vs. the *be-going-to*-future or the *s*-genitive vs. the *of*-construction). This methodology enables us to gain insights on several levels.

First, it allows descriptions of 'alternation' phenomena on a sounder empirical basis and at a finer level of detail than has previously been the case. These descriptions have uses, for example, in applied linguistics, where they may serve to structure teaching materials and modern reference works (in the vein of Biber et al. (1999)). They also have implications for psycholinguistic research – for example in the context of Stallings et al.'s (1998) 'verb disposition hypothesis', which states that individual verbs have dispositions towards certain constructions or constructional processes; our method provides a more sophisticated operationalization of verb disposition than previous studies, in addition to making it possible to (partially) explain lexical bias effects in the first place.

Second, our results bear directly on the analysis of specific, much-discussed phenomena of English grammar, where they generally support analyses based

on constructional semantics and disconfirm analyses based purely on syntactic or pragmatic factors (as in the case of the dative alternation, the passive, and the genitive), or where they show a strong lexical bias of constructional alternatives to exist in addition to previously discussed syntactic or pragmatic factors (as in the case of the verb-particle construction or the two future constructions).

Finally, the results obtained by our methodology have consequences for linguistic theorizing at a more general level; the results show that in some of the paradigm cases of 'alternation', there is clear evidence that each of the two members of the alternating pair is a construction in its own right with its own meaning. These meanings may overlap, thus placing the constructions in a partial paraphrase relationship with each other; still, what is primary are the constructions, and not the paraphrase relation (cf. Goldberg (2002), Gries (2003a: Section 7.1.2), Stefanowitsch (2003) for additional arguments for this position, based on a variety of syntactic, semantic and pragmatic differences between the constructions involved in some of the classic examples of alternating pairs).

Notes

* The order of authors is arbitrary.

1. Note that the method is not restricted to alternating pairs but can be extended to 'alternating triplets' or any other size set of semantically similar constructions. For example, we might extend the set of constructions exemplified in (1) above by adding the *for*-dative or constructions of the type *John {provided/gifted/presented/…} Mary with a book*.

2. Note that we use the Fisher exact p-value (i.e. the statistical probability of error) as a measure of distinctivity. This use of a measure of statistical significance as a measure of association strength might be objected to; usually, an effect size is used to determine association strength, and this effect value is then tested for statistical significance. However, the advantage of the Fisher exact p-value is that it incorporates the size of the effect observed in any particular cross-tabulation, as well as weighing the effect on the basis of the observed frequencies. This sensitivity to frequency seems a desirable property when dealing with natural language data (for example, 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)) (the use of p-values as a measure of association strength is justified in more detail in Stefanowitsch & Gries (2003), following earlier work by Pedersen (1996), Pedersen, Banerjee & Purandare (2003)).

3. A detailed description of the ICE-GB, its design and markup can be found in Greenbaum (1996).

4. The collexemes in Table 3 could be argued to show an additional influence on the choice between active and passive, namely that of genre: note that the top distinctive collexemes

for the passive include a greater number of formal, Latinate verbs than those for the active. This may be due to the uneven distribution of actives and passives across the written and the spoken part of the ICE-GB: in the written part, actives are about 5.5 times more frequent than passives while in the spoken part, actives are about 13 times more frequent than passives. A more refined distinctive-collexeme analysis could take into account the distribution of the constructions across genres, text types, or even the constructions' dispersion across individual samples in the corpus.

5. An interesting problem is exemplified by *have on*: a closer look at the context indicates that the five occurrences in the ICE-GB in fact instantiate two different senses, namely 'wear clothes' (e.g. *You have your bra on* [S1A-090 #252:2:C]) and 'conduct' (e.g. *And they having a bit of a sort of a sale on* [S1A-040 #286:1:A]); while in this case the two senses of *have on* exhibit the same preference for one construction, this need of course not necessarily be the case; different (related or unrelated) senses of the same verb may well exhibit different constructional preferences. Thus, in some cases it might be more precise and rewarding to not just look at the distinctive collexemes of verbs, but of verb *senses*, i.e. verb-sense specific patterns (cf. Hare, McRae & Elman (2003) and recent work by Roland and colleagues, e.g., Roland & Jurafsky (2002)).

6. For those who prefer a more technical support of the line of reasoning just adopted, let us briefly evaluate the proposed methodology in terms of a well-established effectiveness measure, namely F (with precision and recall weighted equally; cf. Manning & Schütze 2000: Section 8.1). Computing all 24 F values (12 constructions predicted according to either our distinctive-collexeme analysis or raw frequency) shows that the average F-value following from the distinctive-collexeme analysis (0.77) is considerably larger ($t = 1.9$; $df = 11$; $p = 0.042$) than that of simply deciding on the basis of the more frequent construction (0.65). Thus, the qualitative argumentation against simply never predicting the less frequent construction is also supported quantitatively by the higher tradeoff of precision and recall resulting from our distinctive-collexeme analysis.

7. Given the degree of predictive power of distinctive collexeme analysis demonstrated here, it is intriguing to think about possible applications. For example, as one reviewer has rightly pointed out, the information yielded by this method might be used to further increase the naturalness of the output of natural language generation systems especially in combination with other context features.

References

Altenberg, B. (1980). Binominal NPs in a thematic perspective: Genitive vs. *of*-constructions in 17th century English. In S. Jacobson (Ed.), *Papers from the Scandinavian Symposium on Syntactic Variation* (pp. 149–172). Stockholm: Almqvist & Wiksell.

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., Johansson, S. Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Bolinger, D. (1975). On the passive in English. In A. Makkai & V. Becker Makkai (Eds.), *The First Lacus Forum 1974* (pp. 57–80). Columbia, SC: Hornbeam Press.

Browman, C. (1986). The hunting of the quark: The particle in English. *Language and Speech, 29*(4), 311–334.

Burt, M.K. (1971). *From Deep to Surface Structure*. New York: Harper Row.

Chomsky, N. A. (1957). *Syntactic Structures*. The Hague: Mouton.

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

Church, K.W., Gale, W., Hanks, P. & Hindle, D. (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.

Deane, P. D. (1987). English possessives, topicality and the Silverstein hierarchy. In J. Aske, N. Beery, L. Michaelis, & H. Filip (Eds.), *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society* (pp. 65–76). Berkeley, CA: Berkeley Linguistics Society.

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

Erteschik-Shir, N. (1979). Discourse constraints on dative movement. In T. Givón (Ed.), *Discourse and Syntax: Syntax and Semantics 12* (pp. 441–487). New York, San Diego & London: Academic Press.

Fillmore, C. J. (1965). *Indirect Object Constructions in English and the Ordering of Transformations*. The Hague: Mouton.

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). Berkeley, CA: Berkeley Linguistics Society.

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

Fisher, C., Gleitman, H., & Gleitman, L. R. (1991). On the semantic content of subcategorization frames. *Cognitive Psychology, 23*(3), 331–392.

Givón, T. (1993). *English Grammar. A Function-based Approach*. Vols 1 & 2. Amsterdam & Philadelphia: John Benjamins.

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

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

Goldberg, A. E. (2002). Surface generalizations: An alternative to alternations. *Cognitive Linguistics, 13*(3), 327–356.

Greenbaum, S. (Ed.) (1996). *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Gries, St. Th. (2002). Evidence in linguistics: Three approaches to genitives in English. In R. M. Brend, W. J. Sullivan & A. R. Lommel (Eds.), *LACUS Forum XXVIII: What constitutes evidence in linguistics?* (pp. 17–31). Fullerton, CA: LACUS.

Gries, St. Th. (2003a). *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. London, New York: Continuum Press.

Gries, St. Th. (2003b). Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics, 1*, 1–28.

Gries, St. Th. (2003c). 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., & Stefanowitsch, A. (forthcoming). Covarying collexemes in the *into*-causative. In M. Achard & S. Kemmer (Eds.), *Language, Culture, and Mind*. Stanford: CSLI Publications.

Gropen, J., Pinker, S., Hollander, M., Goldberg, R., & Wilson, R. (1989). The learnability and acquisition of the dative alternation in English. *Language, 65*(2), 203–257.

Hare, M. L., McRae, K., & Elman, J. L. (2003). Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language, 48*(2), 281–303.

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

Hawkins, R. (1981). Towards an account of the possessive constructions: NP's N and the N of NP. *Journal of Linguistics* 17(2), 247–269.

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Jespersen, O. (1949). *A Modern English Grammar on Historical Principles*, Vol. 7: *Syntax*. Copenhagen: Munksgaard.

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. (1987). *Women, Fire, and Dangerous Things. What Categories Reveal about The Mind*. Chicago & London: The University of Chicago Press.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar*. Vol. 1. *Theoretical Prerequisites*. Stanford: Stanford University Press.

Langacker, R. W. (1991). *Foundations of Cognitive Grammar*. Vol. 2. *Descriptive Application*. Stanford: Stanford University Press.

Langacker, R. W. (1995). Possession and possessive constructions. In J. R. Taylor & R. E. MacLaury (Eds.), *Language and the Cognitive Construal of the World* (pp. 51–79). Berlin & New York: Mouton de Gruyter.

Lees, R. B. (1960). *The Grammar of English Nominalizations*. The Hague: Mouton.

Legum, S. (1968). The verb-particle construction in English: Basic or derived? In B. J. Darden, C.-J. Bailey & A. Davison (Eds.), *Papers from the Fourth Regional Meeting Chicago Linguistic Society* (pp. 50–62). Chicago, IL: Chicago Linguistics Society.

Levin, B. (1993). *English Verb Classes and Alternations. A Preliminary Investigation*. Chicago & London: The University of Chicago Press.

Manning, C. D., & Schütze, H. (2000). *Foundations of Statistical Natural Language Processing*. 4th printing with corrections. Cambridge, MA: The MIT Press.

Murphy, R. (1986). *English Grammar in Use*. 4th revised impression. Cambridge: Cambridge University Press.

Nordquist, D. (1998). *Try and*: a discourse analysis. *Proceedings of the First High Desert Linguistics Society Conference,* April 3–4. Albuquerque, NM.

The Oxford English Dictionary. (1989). Second edition, 20 vols. Oxford: Clarendon Press.

Osselton, N. E. (1988). Thematic genitives. In G. Nixon & J. Honey (Eds.), *An Historic Tongue: Studies in English Linguistics in Memory of Barbara Strang* (pp. 138–144). London: Routledge.

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

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

Pinker, S. (1989). *Learnability and Cognition*: *The Acquisition of Argument Structure*. Cambridge, MA: The MIT Press.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Rice, S. A. (1987). Towards a transitive prototype: Evidence from some atypical English passives. In J. Aske, N. Beery, L. Michaelis & H. Filip (Eds.), *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society* (pp. 422–434). Berkeley, CA: Berkeley Linguistics Society.

Roland, D., & Jurafsky, D. (2002). Verb sense and subcategorization probabilities. In P. Merlo & S. Stevenson (Eds.), *The Lexical Basis of Sentence Processing: Formal, Computational and Experimental Issues* (pp. 303–324). Amsterdam & Philadelphia: John Benjamins.

Stallings, L. M., MacDonald, M. C., & O'Seaghdha, P. G. (1998). Phrasal ordering constraints in sentence production: Phrase length and verb disposition in Heavy-NP Shift. *Journal of Memory and Language, 39*(3), 392–417.

Standwell, G. B. J. (1982). Genitive constructions and functional sentence perspective. *International Review of Applied Linguistics, 20*(4), 257–261.

Stefanowitsch, A. (1998). Possession and partition: The two genitives of English. *Cognitive Linguistics: Explorations, Applications, Research, 23*. Dept. of English, University of Hamburg.

Stefanowitsch, A. (2003). Constructional semantics as a limit to grammatical variation. In G. Rohdenburg & B. Mondorf (Eds.), *Determinants of Grammatical Variation in English* (pp. 155–173). Berlin & New York: Mouton de Gruyter.

Stefanowitsch, A., & Gries, St. Th. (2003). Collostructions: On the interaction between verbs and constructions. *International Journal of Corpus Linguistics, 8*(2), 209–243.

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

Thompson, S. A. (1990). Information flow and dative shift in English discourse. In J. A. Edmondson, C. Feagin & P. Mühlhäusler (Eds.), *Development and Diversity: Language Variation across Time and Space* (pp. 239–253). Dallas: SIL & University of Arlington, TX.

Thompson, S. A., & Koide, Y. (1987). Iconicity and 'indirect objects' in English. *Journal of Pragmatics, 11*(3), 399–406.
Thompson, A. J., & Martinet, A. V. (1986). *A Practical English Grammar*.4th edition. Oxford: Oxford University Press.
Weeber, M., Vos, R., & Baayen, R. H. (2000). Extracting the lowest-frequency words: Pitfalls and possibilities. *Computational Linguistics, 26*(3), 301–317.