Corpus data in usage-based linguistics

What's the right degree of granularity for the analysis of argument structure constructions?*

Stefan Th. Gries
University of California, Santa Barbara

The use of corpus data in cognitive linguistics brings with it a host of methodological problems. One concerns the degree of granularity that provides the most insightful results. The present study investigates two granularity issues – different inflectional forms and (register-)based corpus parts. First, I compare the results of a lemma-based corpus analysis of an English argument structure construction to an inflectional-form-based corpus analysis to determine whether the two approaches result in different suggestions concerning the semantics of the construction at issue. Second, I outline how to determine whether data from different corpus parts/registers result in different semantic generalizations of the same construction and how relevant corpus distinctions can be determined in an objective bottom-up manner.

Keywords: corpus data, granularity, lemmas, registers, ditransitive

1. Introduction

While cognitive linguistics has always been much more concerned with how speakers represent, process, and actually *use* language than many other frameworks in theoretical linguistics, this tendency has gained in importance only in the past few years when the notion of 'usage-based linguistics' has become increasingly frequent in papers and publications. As a result of this development, the number of studies which invoke evidence from actual language usage – corpora – has also increased substantially, a tendency which is not only obvious in cognitive linguistics but also more generally and even in parts of generative linguistics (cf., for a recent example, Kepser and Reis 2005).

* I thank the audiences at the Fourth International Conference on Construction Grammar and Conceptual Structure, Discourse, and Language 2006 for comments as well as John Newman for feedback and discussion. The usual disclaimers apply.

On the one hand, this increase is most welcome since it raises the methodological standards of the discipline of linguistics, which has all too long been plagued by an overly strong reliance on made-up data and judgments concerning these data. On the other hand, the fact that this development is so fairly recent is also responsible for the fact that methodological standards and procedures are still developed, evaluated, and negotiated among those practitioners of the field who apply corpus-based methods in their work. One of the from my point of view most central questions in this regard is concerned with the degree of granularity that provides the most insightful results. This is because corpora provide data on many different levels of hierarchical organization, and not all hierarchical levels are necessarily suited equally well to all tasks.

One case in point is the distinction between lemmas and inflectional word forms. It is probably fair to say that the emphasis of most lexicographers and semanticists has so far been, if only implicitly, on the level of the lemma. For example, in most cognitive-linguistic studies of the semantics of verbs, verbs were discussed by referring to their infinitive form and usually not by addressing the question of whether different inflectional forms exhibit (significantly) different behavior. Similarly, the discussion of how slots of argument structure constructions are filled with verbs and what the verbs filling these slots reveal about the constructions has largely been involving unspecified/infinitive forms. The assumption has been that the semantics of, say, the caused-motion construction is independent of whether the element inserted into the verb slot of the construction is *push* or *pushed*. In particular, the corpus-based methodology to investigate the semantics of argument structure constructions that is currently most fleshed out, collostructional analysis, has so far also used the lemma as the basic unit of analysis, collapsing all the inflectional forms of the words whose occurrences in constructions were investigated (see e.g. Stefanowitsch and Gries 2003, Gries and Stefanowitsch 2004), and so far this approach has resulted in a multitude of findings in different languages (English, German, Dutch, Portuguese, and others) as well as concerning different phenomena (the semantics of constructions, second-language learning, syntactic priming, and others).

However, some recent work (e.g. Rice and Newman 2005; Newman and Rice 2006) has been diverging from this reliance on lemmas, which was rarely ever topicalized explicitly anyway. They argued that the finer resolution of actually inspecting inflectional forms may be more revealing. More specifically, Rice and Newman (2005) discuss how different inflectional forms of several lemmas (e.g. *to think*, *to allow*, and *to rain*) differ in their frequencies both (i) in a complete corpus and (ii) in register-based parts of corpora, concluding that "the frequencies of inflectional forms vary across different register-based parts of corpora and this should be taken into consideration in a corpus-based analyses of linguistic units" (Newman p.c.; cf. also Newman and Rice 2006 as well as Sinclair 1991 for a similar point). In addition, Newman and Rice's (2006) investigation of a sample of the inflectional forms of *to eat* and *to drink* in the British National Corpus (BNC) systematically contrasts spoken and written data to uncover differences between the two modes and implications following from such differences.

Given that the upsurge of corpus-linguistic studies in cognitive linguistics is only a fairly recent development, it comes as no surprise that there are so far very few studies that directly compare the results of the two approaches – lemmas vs. inflectional word forms – in a wider variety of contexts. Also, there is little work that investigates the degree to which register-based parts of corpora result in different cognitive-linguistic analyses (rather than just variational patterns).[1] One obvious exception is the work by members of the research group Quantitative Lexicology and Variational Linguistics. However, as I see it, most of this work seems to be more interested in sociolinguistic and variational issues proper and the methodological implications these have for cognitive linguistics rather than in the conceptual integration of these issues into current cognitive-linguistic or construction grammar theorizing. For example, Heylen's (2004) analysis of word order variation in the German *Mittelfeld* includes 'cognitive-linguistic factors' such as animacy and givenness of referents as well as sociolinguistic variables, but the results are not in turn used to inform a cognitive-linguistic model.

The present study investigates these two granularity issues – lemmas vs inflectional forms and whole corpora vs. register-based corpus parts – on the basis of corpus data from the British Component of the International Corpus of English (ICE-GB). As to the former, I will compare the results of a lemma-based corpus analysis of an English argument structure construction to an inflectional-form-based corpus analysis to determine whether the two approaches result in different suggestions concerning the semantics of the construction in point. As to the latter, I will test whether data from different corpus registers invite different semantic generalizations of the same construction. Obviously, there are many different areas of investigation, of which the present one – the semantics of argument structure constructions – constitutes just one single example. The results of the present study are therefore not intended to once and for all resolve the issue of which levels of granularity to choose, which would require at least a monograph-length treatment in order to investigate the many issues other than argument structure construction semantics. Rather, the present results must therefore be understood as methodological suggestions for what a more comprehensive analysis may ultimately look like and for some initial results for parts of such an analysis.

2. Approximating the semantics of constructions: Collostructional analysis

The method that will be used here is that of collexeme analysis, one method of the family of methods of collostructional analysis mentioned above. Given space constraints, I will introduce the method here only briefly; for more detailed discussion and particularly exemplification the reader is referred to Stefanowitsch and Gries (2003).

1. Another exception is Stefanowitsch and Gries (2008), which will be mentioned briefly below.

Collexeme analysis allows to investigate the semantics of a construction by identifying the words that are associated to a syntactically-defined slot of a construction. It is similar to collocational studies and assumes a Goldbergian kind of Construction Grammar approach to language, according to which lexis and grammar form a continuum of elements. The idea underlying collexeme analysis is that linguistic elements tend to occur in or with other linguistic elements to the degree that they are similar to each other. To investigate the semantics of an argument structure construction, the following steps must be taken:

i. one looks for all examples of the construction (which often involves semi-manual coding);
ii. one retrieves all words/lemmas occurring in a particular syntactic slot of the construction (usually the main verbs) as well as their overall frequencies in the corpus to generate a 2 × 2 co-occurrence table of the kind represented in Table 1 *for each word* (which is referred to as a collexeme); in this table, $a + b$ is the overall frequency of the word/lemma *W* in the corpus, $a + c$ is the overall frequency of the construction *C* in the corpus, *a* is the frequency of co-occurrence of the word and the construction and *N* is the corpus size.
iii. from each such 2 × 2 co-occurrence table, one computes a measure of association strength (called collostruction strength) to determine (a) the *direction* of the co-occurrence, i.e. whether the construction and the word co-occur more or less frequently than expected, and (b) the *strength* of the more-or-less-frequent-than-expected co-occurrence. In most previous studies, the measure of association computed was the logarithm of the *p*-value of a Fisher-Yates exact test to the base of 10, which was multiplied with –1 if the word occurs more often in the construction than expected. This procedure results in high positive values for strongly attracted words, values around 0 for verbs which occur in the construction with chance frequency, and negative values with words that are repelled by the construction. For the lemma *to tell* and the ditransitive in the ICE-GB, for example, Stefanowitsch and Gries obtained the 2×2 co-occurrence table represented as Table 2, the *p*-value for this distribution is 1.596257e–127, and the resulting collostruction strength value of the kind that will be used here is accordingly 126.7969.[2]

Table 1. Schematic 2×2 co-occurrence table for the statistical analysis of collexemes

	Construction *C*	**¬ Construction** *C*	**Row totals**
Word *W*	*a*	*b*	$a + b$
¬ Word *W*	*c*	*d*	$c + d$
Column totals	$a + c$	$b + d$	$a + b + c + d = N$

2. All statistics and graphics were computed and generated with Coll.analysis 3.2a (Gries 2007), a script written in R, an open source programming language and environment for statistical computing (cf. R Development Core Team 2005).

Table 2. 2 × 2 co-occurrence table for the statistical analysis of *to tell* in ditransitives with object NPs in the ICE-GB

	ditransitive	other constructions	Row totals
tell	128 (exp.: 5.93)	666	794
other verbs	907	136,963	137,870
Column totals	1,035	137,629	138,664

Once the analysis of all tables for all the words has been completed, one can then rank the words according to the collostructional strength, and previous work has shown that the top-ranked words – i.e., the words most strongly attracted to a particular construction – provide a multitude of clues about semantic properties of the construction investigated. More specifically, often one can read the semantics of a construction fairly directly off the most strongly attracted collexemes of the construction under investigation. In the following sections, the results of several such analyses for different inflectional forms will be compared to each other.

3. Case study 1: Lemmas vs. inflectional forms in the English ditransitive construction

One of the most thoroughly studied argument structure constructions is the English ditransitive construction, which is exemplified in (1).

(1) a. *He gave her the book.*
b. *She told him a story.*
c. $\text{SUBJ}_{\text{AGENT}}$ V OBJ_{REC} $\text{OBJ}_{\text{THEME}}$

Given that much of the semantics of this construction is so well-known (cf. Goldberg 1995: ch. 6 for one authoritative account), it provides an ideal test case against which the results of different methodological procedures can be evaluated. One first and simple way of evaluating different levels of granularity would be to compare the results of a collexeme analysis based on verb lemmas occurring in the English ditransitive to those collexeme analyses based on individual inflectional forms of the verbs occurring in the English ditransitive. The central aspects to be singled out for comparison are the degrees to which

- the senses postulated in analyses of the ditransitive are reflected uniformly across the collexeme analyses;
- the ranking of the verbs in the lemma-based analysis is correlated with the ranking of the verbs if only particular inflectional forms are included into the analysis.

Results of a lemma-based collexeme analysis of the English ditransitive have already been published (cf. Stefanowitsch and Gries 2003: 229), but in this chapter I will use a more comprehensive data set. Stefanowitsch and Gries restricted their analysis to ditransitives with noun objects while I will use all ditransitives. The results, computed as discussed in more detail above are summarized in Table 3. It is easy to see that both the sense of transfer that is usually associated with this construction and the many semantic extensions the ditransitive is argued to have are strongly reflected in the top collexemes: transfer (*give*, *send*, *lend*, *award*), the satisfaction-condition-imply-transfer sense (*offer*, *promise*, *owe*, *guarantee*), the enabling-transfer sense (*allow*), the causing-not-to-receive sense (*deny*), the future-transfer sense (*grant*), the extended communication senses (*tell*, *ask*, *teach*) etc. The verbs that are most strongly repelled are the fairly high frequency verbs *make*, *do*, *find*, *call*, *get*, and *take*, whose semantic characteristics have little to do with what has usually been considered central to the ditransitive. (Note that while *get* in the ditransitive is of course associated with a 'change of possession' sense, it is much more strongly associated with transfer in the transitive construction with its different order of coarse semantic roles: In *I got some dried flowers in vases*, where the subject is the recipient and not the agent. [3])

Table 3. Top thirty collexemes of the ditransitive construction in the ICE-GB

Verb	CollStr	Rank/no of types	Verb	CollStr	Rank/no of types
give	infinity	1	*grant*	10.59	0.82
tell	infinity	0.99	*warn*	10.24	0.81
ask	73.08	0.98	*award*	9.21	0.8
send	71.88	0.96	*persuade*	8.09	0.79
show	55.73	0.95	*allow*	7.7	0.78
offer	52.42	0.94	*guarantee*	7.37	0.76
convince	36.09	0.93	*deny*	6.52	0.75
cost	26.35	0.92	*earn*	6.2	0.74
inform	23.29	0.91	*pay*	4.85	0.73
teach	22.41	0.89	*allocate*	4.6	0.72
assure	20.16	0.88	*accord*	4.38	0.71
remind	19.36	0.87	*buy*	4.3	0.69
lend	14.62	0.86	*assign*	4.24	0.68
promise	12.65	0.85	*advise*	3.77	0.67
owe	10.77	0.84	*wish*	3.66	0.66

3. A look at WordNet 2.0 strongly supports this point: The transitive change-of-possession sense is by far the most frequent one of *get* while the ditransitive use is only the sixth most frequent one, with only 16 percent of the occurrences of the transitive one.

In order to compare these results to those of analyses based on inflectional forms, however, one intermediate step is necessary. The measure CollStr is influenced by the sample size. Thus, when different analyses are compared, one cannot directly compare the CollStr values since the differences one finds may just be a function of the different sample sizes. Also, while so far virtually all collexeme analyses have nearly always been based only on the ranks of the collexemes rather than on the absolute value of Collstr (i.e. 73.08 for *ask*), it would not be optimal to use the ranks of the verbs because differently sized samples will also result in different ranges of ranks, which make a straightforward comparison difficult. In this study, I will therefore use $^{\text{rank of type}}/_{\text{no of types}}$ of each verb (where ranks are assigned in ascending order, i.e., *give* gets the highest rank) since this way all ranks for verbs fall into the interval 0...1 with most strongly attracted verbs scoring near 1 and the least attracted and repelled near 0.

As the next step, I performed separate collexeme analyses for the following five (classes of) inflectional forms as annotated in the ICE-GB (infinitives, *ing*-participle, past participle, past tense, present tense). That is, in each of these analyses, the marginal totals accordingly only took verbs with the particular inflectional ending into consideration. I then ranked the verbs according to their CollStr values and computed rank of type/no of types. Finally, I plotted the ranks obtained for all inflectional verb forms against the ranks of all lemmas; if verbs did not occur in a particular verb form, their rank was set to 0 and they will be displayed rotated by 90° in the plots below. Finally, I added a linear regression line (dotted), the main diagonal (solid) and a locally weighted robust regression (curved solid) to determine which verbs deviate (how much) from from the overall pattern. The logic of the approach is to determine on the basis of the graphs whether the analysis based on inflectional forms results in different semantic classes than the lemma-based analysis.

Beginning with some general results, it can be noted that the overall fit between the ranks of the verb lemmas and the ranks of the same verbs' inflectional forms is quite good: adjusted R^2 is always highly significant and with the exception of present tense always higher than 65%. Second, across all graphs there is a relatively clear pattern such that the fit of the linear regression lines is best at the extremes: the verbs in the bottom left corner and the top right corner usually appear linearly ordered whereas the fit is worse in the middle, mostly with a tendency for the lemma rank to be higher than the inflectional form's rank. Third, note that some graphs hardly feature any verbs above the main diagonal (esp., Figure 2 and Figure 4), indicating that the inflectional form-based analysis did not rank verbs highly, which the lemma analysis would not, too. Finally, note that the patterning of the verbs that occur in the lemma ranking but not in the inflectional ranking is largely unsystematic: They are from the whole range of ranks for the lemma so, unlike what one might have suspected, the verbs that are lacking particular tense forms do not form a homogeneous group on the level of the lemma, which is good since this could have indicated some bias (due to frequencies etc.).

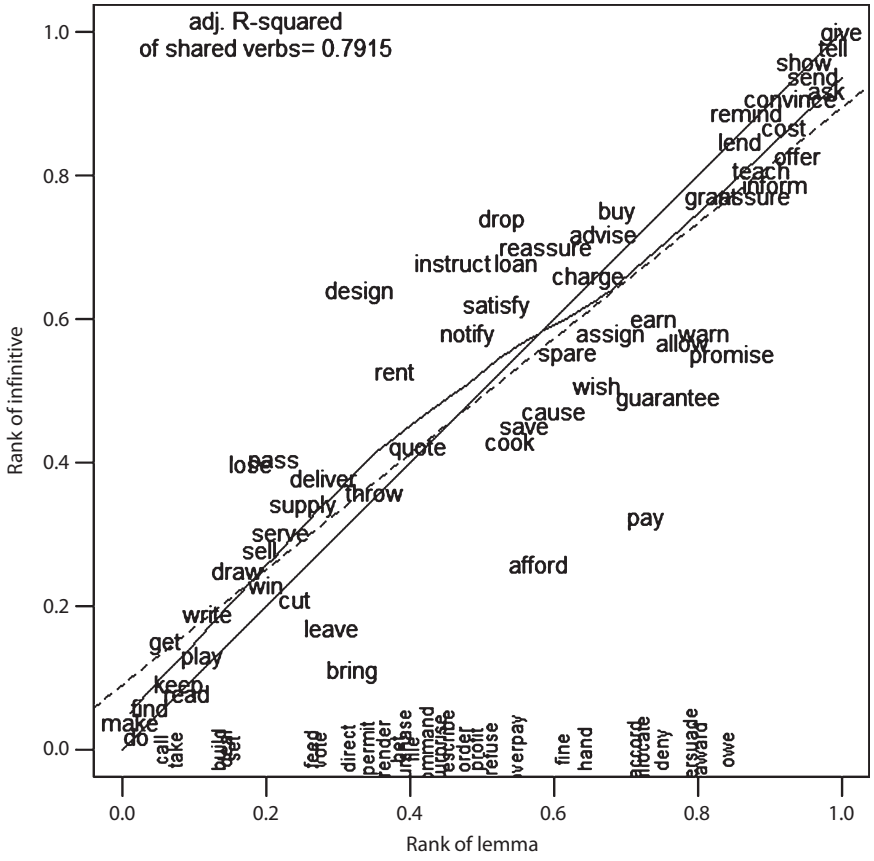

Figure 1. CollStr ranks of infinitives vs. CollStr ranks of lemmas in the English ditransitive

Looking more specifically at the individual graphs shows that the overall conclusions about the semantics of the ditransitive remain fairly constant across all analyses. The two verbs that are most strongly attracted to the ditransitive in the overall lemma analysis – *give* and *tell* – are also attracted most strongly in the case of the individual verb forms (their order changes twice, though). Also, even though there is some variation among the verbs that immediately follow these two in the list of highly attracted verbs, most of them fit the semantics of transfer and its metaphorical extensions (in particular communication) very well: *offer*, *show*, *send*, and *ask* are among the top ten in nearly every single collexeme analysis. In a similar vein, the verbs in the bottom left corner are also relatively homogeneous, comprising mostly high-frequency verbs that are much less specific and less revealing as far as the ditransitive's semantics are concerned.

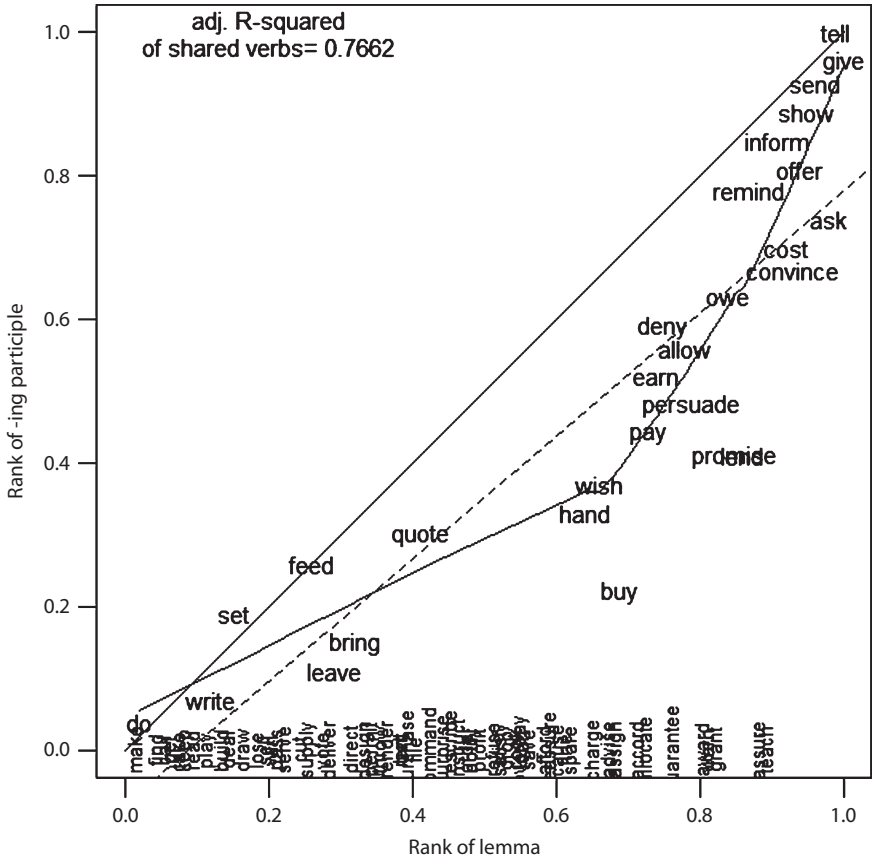

Figure 2. CollStr ranks of ing-participles vs. CollStr ranks of lemmas in the English ditransitive

If we now look at the graphs separately, we find much more variation. So for example, the regression lines for the infinitive and many of the verbs are extremely close to the main diagonal, indicating a particularly good fit. Some verbs which are fairly much below the regression lines are *afford*, *bring*, *guarantee*, *promise,* while verbs such as *design*, *drop*, *instruct*, and *reassure* are among the verbs that are highest above the regression lines and the main diagonal. On the one hand, these findings suggest that there are in fact differences between the distributions of the lemmas and the infinitive forms – otherwise all words and the regression lines would be on the main diagonal.

On the other hand, however, the most important thing to note is that this patterning does not at all change the semantic interpretation of the construction for two reasons: In most cases, the verbs exhibiting marked differences between the lemma-based analysis and the verb form-based analysis

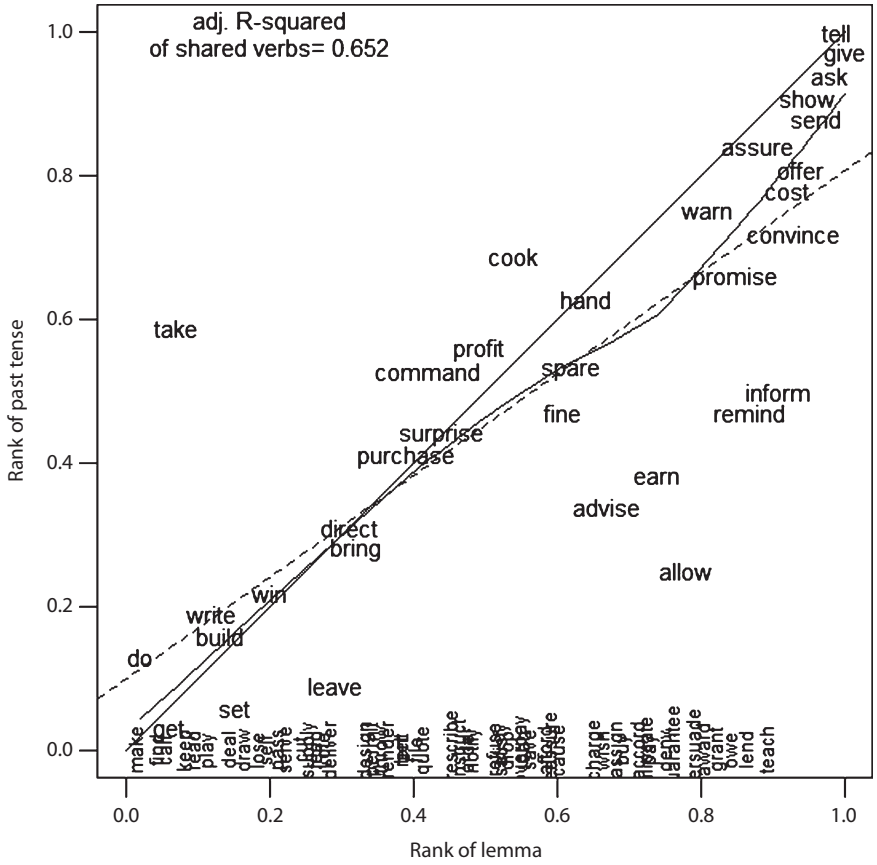

Figure 3. CollStr ranks of past tenses vs. CollStr ranks of lemmas in the English ditransitive

- are in fact not associated with the meaning of transfer at all, or
- represent semantic verb classes which are already represented by verbs which (i) behave identical in both analyses and (ii) are higher-ranking – i.e. more in the top right corner – anyway.[4]

As to the former, some of the outlier verbs whose infinitive distribution differs from the lemma distribution that only take a second object because that is contributed by the ditransitive construction are *design* and *drop*. As to the latter, outlier verbs such as *instruct* and *reassure*, by contrast, belong to semantic classes that are already represented by verbs that are even more strongly associated to the ditransitive anyway (such as, in this case, communication verbs like *tell*, *ask*, *convince*, and even *assure*). Also, the

4. This is of course only an argument if – as in this study – the main interest is on identifying constructional semantics. On other occasions, this kind of difference might be revealing.

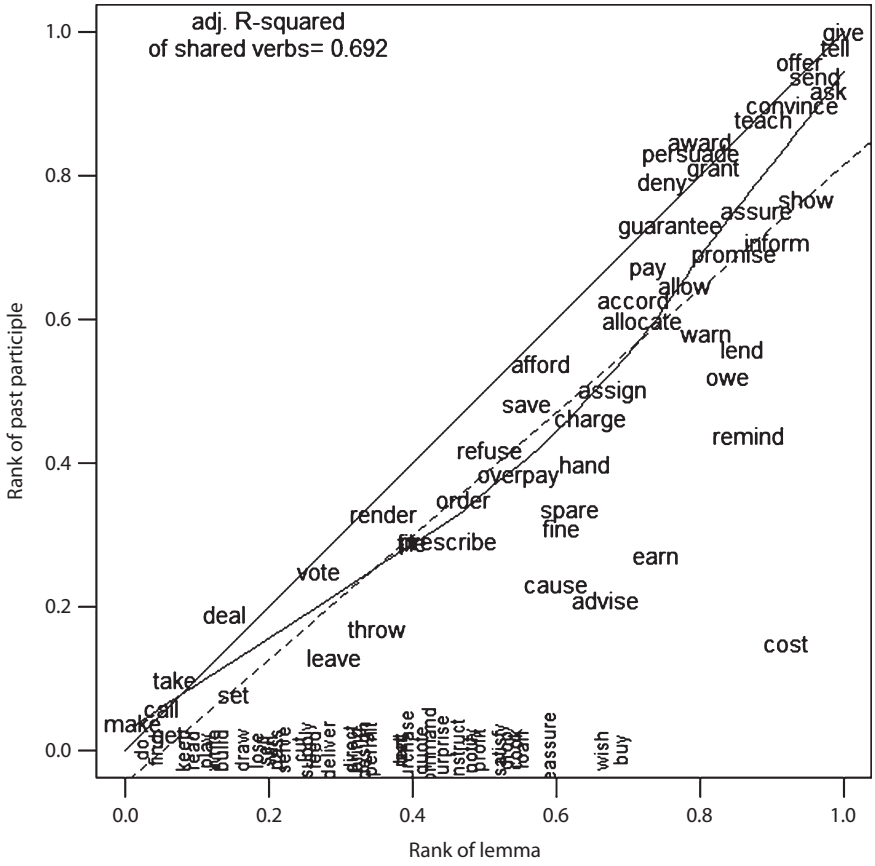

Figure 4. CollStr ranks of past participles vs. CollStr ranks of lemmas in the English ditransitive

satisfaction-condition class verb exemplified by the outlier verbs *promise* and *guarantee* is already represented by *owe* and *grant*, which score nearly identically on both axes. Even the commercial transaction frame represented by the outlier verbs *afford* and *pay* is already represented by higher-ranked verbs such as *buy*, *charge*, and maybe *cost*. All in all, there are differences between the number of verb lemmas and their rankings on the one hand and the number of infinitives and their rankings on the other hand, but these differences are practically meaningless since the semantic classes whose infinitives make a difference are either represented by higher-ranking collexemes of the same class anyway or are actually false hits, i.e. verbs outputted by the infinitive analysis but which in a semantic analysis would not rank equally highly in the first place.

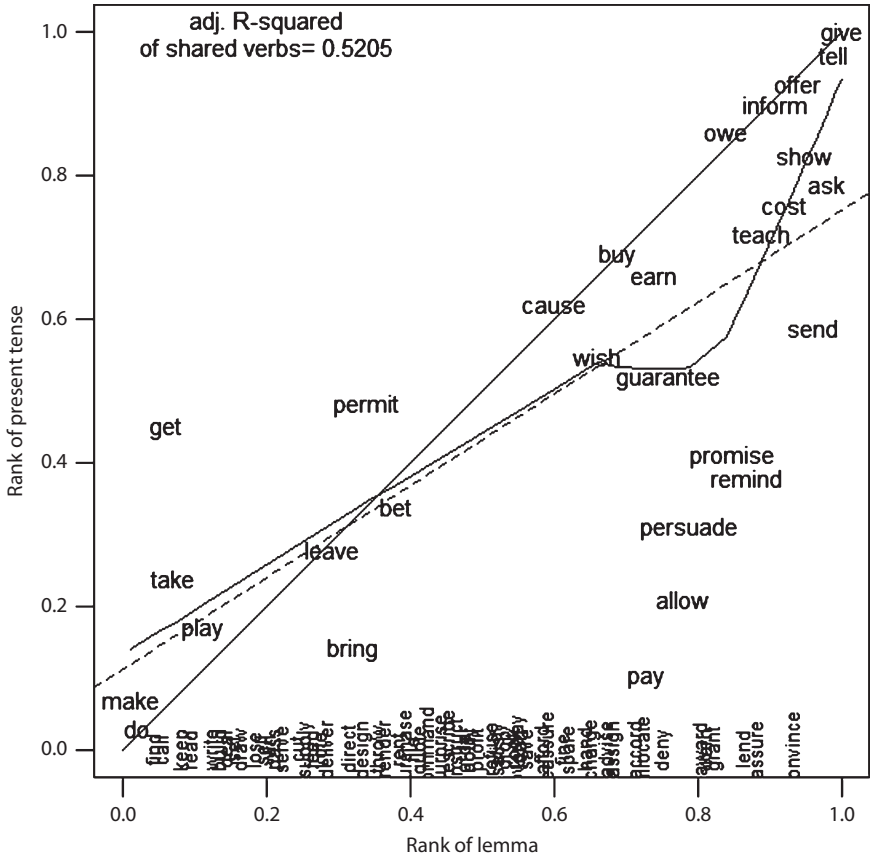

Figure 5. CollStr ranks of present tenses vs. CollStr ranks of lemmas in the English ditransitive

Similar results are obtained for some of the other graphs. For the past tense, *take* and *cook* are verbs whose reading of transfer to an additional person is only contributed by the ditransitive construction. In addition, the outlier verbs *inform*, *remind*, and *advise* are all communication verbs, a class already instantiated by *tell* and *ask*, the class represented by *earn* is established by *profit* and maybe *cost*, and the only verb which does not behave about equally in both analyses and is not a member of an otherwise straightforwardly represented class is *allow*. Just about the same arguments apply to the past participle results, with the exception that the noteworthy outlier of *cost* must be attributed to the absence of *cost* in the passive, which, however, is also not something meaningful about the ditransitive. The results for the *ing*-participle look slightly different because the verbs are more below the main diagonal and because the smoothed regression line has more curvature. However, the results are still essentially similar because most verbs are grouped around the straight regression, and the three which are not

again belong to groups represented by other verbs among the top ten (the class of *buy* is represented by *cost*, *pay*, and *earn*, *promise* by *owe*, and *lend* by *give* and *send*).

The only slightly more striking exception to the above clear patterning are the results for the present tense. While many verbs are perfectly on the main diagonal, the regressions' fits are bad, the curvature of the smoothed regression line is even more extreme than with the *-ing* participle and the verbs are much scattered in the graph. While the class of communication verbs instantiated by *persuade* and *remind* is represented by *tell*, *ask*, and *teach*, the latter verbs are different in their subcategorization behavior. Also, while the *pay* class is reflected by *buy* and *cost*, as is the *promise* class by *guarantee* and maybe *allow* by *permit*, the distances between the verbs and the rank of the verbs which do behave similarly in both analyses are just much larger than in all other graphs. Thus, most semantic relations are represented in both analyses, but the relation is more tenuous than in the results for the other inflectional forms.

All in all, however, the picture is fairly clear. The more fine-grained analysis of the ditransitive using inflectional forms changes some of the *quantitative* results, but the *qualitative* changes are usually minor in the sense that all regressions were highly significant and provided good fits, and the semantic conclusions invited by the most strongly represented collexemes are very much the same. On the basis of the present data, therefore, there is no need either for the finer resolution of an analysis based on inflectional forms or a more comprehensive lemma-based analysis. These results are compatible with studies in other areas. For example, Pickering and Branigan (1998) find that different morphological forms of verbs in syntactic priming studies do not result in differently strong priming effects.

The following section will discuss methods to whether a division of the corpus into different register-based parts yields results different from the overall analysis. While space precludes an actual, full-fledged analysis of the data, I will briefly outline a method of how these data could be utilized.

4. Case study 2: Registers and the English ditransitive construction

While the previous section has been concerned with a level of granularity that was defined on the basis of the actual linguistic forms that are being investigated, other levels of granularity – i.e. other ways of dividing up the data – are more concerned with quasi extra-linguistic aspects of the data. The probably most widespread division is that of spoken vs. written data, and this is also one which Newman and Rice consider in their work and to which they attribute some importance.

This is of course a laudable approach because it allows the researcher to inspect the data with a, as it were, higher resolution: For example, data that may seem overall heterogeneous may in fact be very homogeneous when the modes spoken and written are looked at individually. As a matter of fact, however, Stefanowitsch and Gries (2008) investigate to what degree the difference between speaking and writing influences

distributional patterns in active vs. passive voice, verb-particle constructions, and *will*-future vs. *going-to*-future. They conclude that

> ... there is no evidence so far to suggest that constructional semantics [...] interacts with register/channel [their term for the distinction of speaking vs. writing; STG] in such a way that there are differences in a construction's meaning across register classes. (149)

Thus, while making a distinction between speaking and writing is laudable in principle, there is as yet little support that this distinction yields more than just numerically different results. However, there is also a more fundamental problem: As Gries (2006) argues at length, investigating spoken vs. written data is in a way just a wild guess at the level where one hopes to find the truly revealing patterns or the largest and, therefore, most important differences in one's data. However, without a comparative investigation of (i) which levels of granularity exhibit the largest differences in their patterning and (ii) which effects at this particular level are largest, it remains unclear whether the effects found between spoken and written data are in fact as noteworthy as they seem. To give an example, Gries (2006) investigates the frequency of the present perfect in different corpora of English and contrasts the results with his own results concerning the present perfect in the ICE-GB. To exemplify exactly the problem just raised, he then goes on to compare the different frequencies of present perfects in spoken data and in written data with the different frequencies of present perfects in written printed data and in written nonprinted data. His most interesting finding in this connection, however, is that although the contrast 'spoken vs. written' may appear to be the more fundamental one – after all, 'written printed vs. written nonprinted' is 'only' a within-mode distinction and, thus, lower in a taxonomy of corpus distinctions – it turns out that the difference between the present perfect frequencies within writing is actually slightly larger than between speaking and writing, which is why on the basis of the data alone, an analyst should attend more to the latter contrast than the former. If we generalize from that one particular example (cf. Gries 2006 for more comprehensive discussion), it would be more useful to test several divisions of the corpus and then decide on the basis of the results – not on the basis of any researcher's preconceptions – which level and which effects to attend to. (This is by no means to imply that the distinction 'spoken vs. written' is an unreasonable one, quite the contrary. My point is just that one does not know *beforehand* whether it is the *most* useful one, the one to start with, or the one to focus on most, which is why the above kind of *bottom-up* identification of relevant distinctions is ultimately more rewarding.)

The present study will therefore test the relevance of the kind of *a priori* distinctions such as spoken vs. written just as Stefanowitsch and Gries (2008) but at the same time go beyond that study by also testing whether this distinction is in fact relevant. The overall logic underlying the present approach is similar to that of Gries (2006). As before, I extracted all ditransitives from the ICE-GB as well as the frequencies of all

verbs that occur at least once in the ditransitive in the construction and in the corpus. In addition, for each ditransitive construction, I stored

- the mode in which it occurs: spoken vs. written;
- the register in which it occurs: {spoken dialog} vs. {spoken monolog} vs. {spoken mix} vs. {written printed} vs. {written nonprinted};
- the sub-register in which it occurs: {spoken dialog private} vs. {spoken dialog public} vs. {spoken monolog scripted} vs. {spoken monolog unscripted} vs. {spoken broadcast mix} vs. {written printed academic} vs. {written printed creative} vs. {written printed instructional} vs. {written printed nonacademic} vs. {written printed persuasive} vs. {written printed reportage} vs. {written nonprinted letters} vs. {nonprofessional}.

Then, I wrote R scripts that computed two collexeme analyses for the modes, five collexeme analyses for the registers, and thirteen collexeme analyses for the sub-registers; for ease of comparability, these were all lemma-based analyses. For each of these analyses, the resulting CollStr values were then transformed into the above kind of rank values ranging from 1 (for the most strongly attracted verb) to 0 (for the most strongly repelled verb).

The most straightforward way of using the data would be the same as above. One can plot different modes, registers, or sub-registers against each other to again determine whether all semantic classes obtained in one analysis are also represented in another analysis. As a first example, let us look here only at the distinction invoked by Newman and Rice, spoken vs. written. Figure 6 plots the ranks of the verb lemmas in speaking against the results of the analysis for the verb lemmas in writing.

As is obvious, the correlation between the spoken data and the written data is rather weak: Adjusted R^2 is fairly small especially when compared to the results reported above and the verbs are widely scattered throughout the graph rather than being close to either the main diagonal or the regression line. The main conclusion following from this result is somewhat ambiguous, though. On the one hand, it is obvious that, compared to different inflectional forms, the difference between speaking and writing is huge. Put differently, before an analyst devoted any time to exploring variability of inflectional forms of the verbs in the ditransitive, an analysis of the different modes offers much more variability to account for. On the other hand, however, it is equally obvious that in this case again the results do not change the overall interpretation of the analysis: As before, *give* and *tell* are at the top of the list (in both modes), underscoring the centrality of transfer and communication as transfer for this construction. Also, most major sense extensions are again represented by verbs that are ranked highly both in speaking and writing. Thus, a low correlation between the different kinds of corpus data is only a necessary, but not a sufficient, condition for distinctions that are worth exploring in more detail.

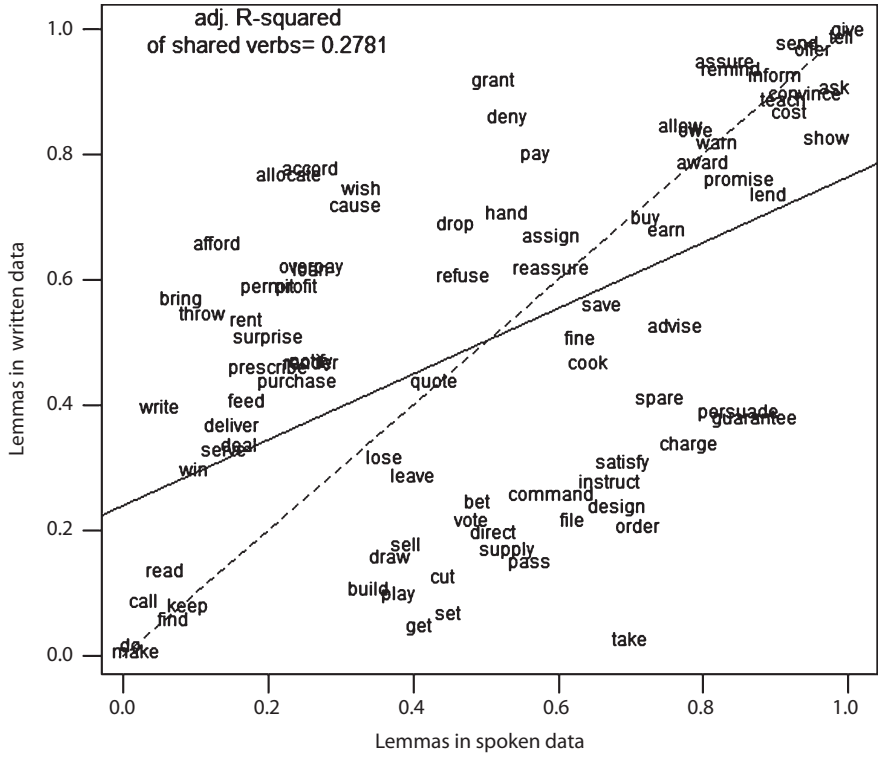

Figure 6 CollStr ranks of verb lemmas in the spoken vs. in the written data

Unfortunately, this approach does not yet address the issue raised at the beginning of this section since again what is tested is just the preconception that spoken vs. written may be a relevant distinction, which it did not turn out to be. However, the discussion of the results already points to a strategy to overcome this shortcoming. If one is interested in the way different modes or registers influence your results, the natural extension of this method would be to test several potentially relevant distinctions at the same time and then check where the largest variation arises (i.e. small R^2's). These cases with a lot of variation are then scrutinized using graphical displays like the above to determine whether they license varying semantic conclusions one would want to account for. An alternative, and probably even better, approach would be to combine the data of the separate collexeme analyses into one table and then let 'the data decide' which groups are manifested, which would render the choice of which distinctions to investigate objective. In this particular case, a technique that can be applied straightforwardly is a principal components analysis (PCA). A PCA is an exploratory data analysis technique which tries to compress the data in a table with *n* columns by grouping columns that behave highly similarly into mutually orthogonal (i.e. independent) principal components. The two theoretically possible extreme outcomes would be (i) a

compression into a single principal component because all *n* columns are very highly intercorrelated, or (ii) the columns are completely uncorrelated so that the data cannot be compressed at all. Thus, the number of principal components remains *n* and each of it can only represent what was one column before the analysis. From this latter extreme, one common criterion for deciding on the number of principal components follows: One usually retains those principal components that can account for more than one variable.[5]

To exemplify the approach, I will briefly report on the results of a PCA on the basis of the CollStr values for the lemma analyses of the whole corpus, of the five registers and of the twelve sub-registers.[6] In this particular case, there is a relatively clear solution. The PCA identifies four principal components with *Eigenvalues* larger than 1; these four principal components can account for 72.35% of the variance of all 18 original variables. To determine, however, what the principal components mean, one can now turn to the corpus parts which load highest on them (I restrict myself to *Eigenvalues* larger than 0.55 because these allow for the most comprehensive and at the same time mutually most exclusive coverage of all columns):

- PC_1: {spoken monolog}, {spoken monolog scripted}, {spoken monolog unscripted}, {spoken dialog public}, {whole corpus};
- PC_2: {written printed}, {written printed persuasive}, {written printed nonacademic}, {written printed academic}, {written printed nonprofessional}, {written printed instructional};
- PC_3: {written nonprinted}, {written nonprinted letters}, {written printed creative};
- PC_4: {spoken dialog}, {spoken dialog private}.[7]

If one now tries to interpret the components on the basis of the corpus parts on which they load highly, then the first component reflects the spoken part of the corpus without private dialog. The second component comprises exclusively written printed data; the third component has the two written unprinted corpus parts plus creative writing; the final component is spoken private dialog. Now, if a linguist assumes a usage-based analysis of the ditransitive that should take into consideration different registers, then these are the register-defined corpus groupings that the linguist should investigate: This is because these four groups of corpus parts are exactly the parts which are most homogeneous internally and most heterogeneous externally with respect to how verbs are attracted to the ditransitive construction. Note also that this distinction is not

5. Technically, this criterion is represented using so-called *Eigenvalues*, which specify the amount of variables a principal component can explain. Cf. Gries (2006) for an alternative methodology using a hierarchical cluster analysis as well as an approach using PCA for measuring corpus homogeneity.

6. The sub-register {spoken mix broadcast} was omitted from analysis because it is the only sub-register within {spoken mix} and thus redundant.

7. The two sub-registers not covered in the above list, {written printed reportage} and {spoken mix broadcast}, load highest on PC_1 and PC_2.

simply the one between speaking and writing (as utilized by Newman and Rice), but also not one just within the five registers or within the thirteen sub-registers – rather, it cuts across *all three levels* of corpus categorization. The same conclusion may actually be arrived at for the difference between Belgian and Netherlandic Dutch that figures prominently in the work of the above-mentioned research unit QLVL. While the factor variety may yield significant results – as it does – it is unclear whether other divisions of the corpus may not result in much more pronounced differences: only an exploratory bottom-up analysis of the kind advocated here can answer this question. The most important lesson to learn from this is that the distinctions one brings to the data as an analyst *a priori* need not at all coincide with the largest differences in the data, those that are actually reflected in the data, or those that are most noteworthy or theoretically revealing. The following section will bring together all findings and conclude.

5. Conclusions

This study started out from the fact that different analyses in usage-based linguistics have been based on different levels of granularity: Some studies were based on complete corpora (or at least parts of corpora that were not further distinguished) while others emphasized a spoken-vs.-written distinction; some studies looked at lemmas whereas others focus on different inflectional forms.

On the basis of the present data, I hope to have shown two main things. First, not all distinctions that are meaningful from a linguistic point of view result in relevant meaningful differences. True, the distinctions tested here result in quantitative changes, but it was shown for both linguistic distinctions (inflectional forms) and situationally defined text types (registers) that these quantitative differences need not result in qualitative interpretive differences of interest. This finding supports the results by Stefanowitsch and Gries (2008) and is also reminiscent of the main lesson one could draw from the lively discussion of how to construct and constrain polysemy networks triggered by Sandra and Rice (1995), who argued – convincingly, I think – that analyses often tended to posit distinctions many of which made sense to linguists but may in fact not have been relevant to speakers' actual linguistic systems. Again, note that this does by no means imply that the distinctions made by other scholars (inflectional forms, registers, varieties, ...) are irrelevant – it just means that (i) they aim at only one level of granularity and (ii) only at one set of factor levels at that level of generality. All I am saying here is that a more comprehensive approach may often be more revealing.

Second, for those usage-based linguists who suspect that register or genre differences are relevant to one's phenomenon under investigation I introduced a method to infer in a bottom-up fashion the corpus parts that exhibit the most pronounced and thus probably most relevant differences in patterning; the major advantages of the method are that (i) it can cut across different levels of categorization – something linguists are often not willing to do – and (ii) the identification of the relevant corpus

parts is done completely objectively. Thus, since the method is based on detecting differences in a bottom-up manner, this method can narrow down the search space for relevant corpus distinctions considerably while at the same time avoiding an inflation of distinctions of dubious relevance to the actual speaker's linguistic system. Note in passing how this method is compatible with Langacker's concern with the "hierarchy of low-level structures [...] that specify the actual array of subcases and specific instances that support and give rise to the higher-level generalization" (Langacker 1991: 281f.). Given these characteristics, this method should actually be extremely relevant to all linguists who regard themselves as usage-based linguists – if the distinctions responsible for differences in one's analysis are directly derived from actual usage data, how much more usage-based can one get?

Of course, while I feel that the scope and applicability of the method as such is enormous, I am the first to admit that more extended testing of phenomena other than argument structure semantics is necessary, as may be refinements and extensions. In addition, given the growing recognition of the relevance of usage data, the exploration of how splitting corpus data along the above lines and/or resampling approaches may open up a variety of new perspectives on how usage data can be exploited fruitfully.

Finally, I should like to point out that the methods proposed above require neither much data beyond what most corpus-linguistic methods already provide nor a huge set of software applications. In fact, all of the above has been performed with data that are usually available anyway since, for example, when one retrieves all instances of the ditransitive from the ICE-GB, each hit comes with the file name, and thus the register etc., anyway. Also, the only software necessary to do the retrieval as well as all computations and graphics is R. I therefore hope that the present work will stimulate future studies exploring the largely uncharted issues of granularity and corpus parts in usage-based cognitive linguistics.

References

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

Gries, Stefan Th. 2006. Exploring variability within and between corpora: some methodological considerations. *Corpora* 1.2: 109–151.

—— 2007. *Coll.analysis 3.2a.* A script for R.

—— & Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspectives on 'alternations.' *International Journal of Corpus Linguistics* 9.1: 97–129.

Heylen, Kris. 2004. Methodological issues in usage-based linguistics. Paper presented at Current Trends in Cognitive Linguistics, Hamburg, Germany.

Kepser, Stephan & Marga Reis, eds. 2005. *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*. Berlin, Heidelberg & New York: Mouton de Gruyter.

Langacker, Ronald W. 1991. *Concept, Image, and Symbol: The Cognitive Basis of Grammar*. Berlin & New York: Mouton de Gruyter.

Newman, John & Sally Rice. 2006. Transitivity schemas of English EAT and DRINK in the BNC. In St. Th. Gries, & A. Stefanowitsch, eds., Corpora in Cognitive Linguistics: *Corpus-Based Approaches to Syntax and Lexis*, 225–260. Berlin & New York: Mouton de Gruyter.

—— & —— 2006. English adjectival inflection: a *radical* Radical Construction Grammar approach. Paper presented at Conceptual Structure, Discourse, and Language, San Diego, CA, USA.

Pickering, Martin J. & Holly P. Branigan. 1998. The representation of verbs: evidence from syntactic priming in language production. *Journal of Memory and Language* 39.4: 633–651.

R Development Core Team. 2005. *R 2.2.1 – A Language And Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. ISBN 3–900051–07–0, <http://www.R-project.org>.

Rice, Sally & John Newman. 2005. Inflectional islands. Paper presented at the International Cognitive Linguistics Conference, Seoul, South Korea.

Sandra, Dominiek & Sally Rice. 1995. Network analyses of prepositional meaning: Mirroring whose mind – the linguist's or the language user's? *Cognitive Linguistics* 6.1: 89–130.

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

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8.2: 20–43.

—— & —— 2008. Channel and constructional meaning: a collostructional case study. In G. Kristiansen & and R. Dirven, eds., *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*, 129–152. Berlin & New York: Mouton de Gruyter.