4

Cluster Analysis and the Identification of Collexeme Classes

STEFAN TH. GRIES AND ANATOL STEFANOWITSCH[*]

1 Introduction

In recent work, we have developed a family of collocational methods geared specifically towards investigating the relationship between words and constructions, referred to as *collostructional analysis*. While this method is rigorously quantitative and objective with respect to the way in which it identifies the strength and direction of association between a construction and the words occurring in this construction, it still relies on qualitative and subjective arguments concerning the way in which the results are interpreted. In particular, once we have identified a set of words that are significantly associated with a given (slot of a) construction, we typically group these into semantic classes based on introspection and common-sense arguments.

In this paper, we address this problem and sketch out a potential solution. We investigate to what degree cluster analytic techniques can identify semantic classes more objectively and precisely, and thus shed light on the most prototypical sense(s) of a construction as well as subsenses instantiated by coherent semantic classes of words occurring in it. In the remainder of this Section, we discuss the problem and its potential solution

[*] Parts of this research were presented at the ICCG-3 and the CSDL-7; the authors would like to thank the participants for discussion.

Empirical and Experimental Methods in Cognitive/Functional Research.
Sally Rice and John Newman (eds.).

in more detail in the context of our previous research. In Section 2, we present three case studies in which cluster analysis is applied to three different constructions.

1.1 Previous Work on Collostructional Analysis

The set of methods which we refer to as collostructional analysis (Stefanowitsch & Gries 2003, 2005; Gries & Stefanowitsch 2004a, b) is grounded in two frameworks, one theoretical and one methodological. The theoretical framework is provided by constructional theories of grammar, i.e., theories which view grammatical structures of various degrees of complexity and schematicity (morphemes, words, partially lexically filled idioms, argument structure constructions etc.) as linguistic signs, i.e., as meaningful linguistic elements in their own right. Collostructional analysis is usually phrased in the terminology of the cognitively-oriented version of Construction Grammar propagated, for example, by Lakoff (1987) and Goldberg (1995), but it is not dependent on this theory specifically—the methods introduced in our earlier work and below are useful in any framework that takes grammatical structures to be meaningful.

The methodological framework is that of quantitative corpus linguistics. This approach is characterized by three features that distinguish it slightly from both traditional corpus linguistics and computational linguistics: first, wherever possible it is based on naturally-occurring language data from representative and balanced corpora;[1] second, the linguistic phenomenon in question is retrieved exhaustively from the corpus (i.e., with maximal precision and recall, which usually requires manual post-editing of many thousands of hits); and third, the data are subjected to strict quantification and statistical evaluation; cf. Stefanowitsch & Gries (2009) for discussion of how this approach differs from some of the British school of corpus linguistics).

Collostructional analysis comprises three different though related methods. The first of these is *collexeme analysis*: this method identifies the relation between a construction (C) and the words $W_{1\text{-}n}$ occurring in a given slot of C (extending previous work on collocation strength, cf. especially Church, Gale, Hanks, & Hindle 1991). Like all (or most) collocational methods, it is based on a two-by-two table of cooccurrence frequencies as

[1] Strictly speaking, it is of course impossible to define whether a particular corpus is representative of some medium/channel, variety, register, or language as a whole since it is not possible to describe the population of linguistic elements from which a corpus constitutes a mere sample – 'representative and balanced' is therefore intended to mean that no particular medium/channel, variety, or register is a priori unduly over- or underrepresented in the corpus. Thus, the British National Corpus and the British Component of the International Corpus of English, for example, are considered as belonging to this category.

shown schematically in Table 1; frequencies in bold type are obtained from the corpus directly while the remaining ones result from subtractions.

	Construction C	Other Constructions	Row Totals
Word W_x	**a**	b	**a+b**
Other Words	c	d	c+d
Column Totals	**a+c**	b+d	**N=a+b+c+d**

Table 1. Cooccurrence table for a collexeme analysis

The figure of interest here is *a*, the cooccurrence frequency of word Wx in the construction C under investigation. If the observed frequency of *a* is significantly higher or lower than expected, the relation between W and C is one of attraction or repulsion respectively (W is then said to be a significantly attracted or repelled *collexeme* of C, hence the name of the method). Any distributional statistic can in theory be used to quantify the strength of the attraction or repulsion. For reasons discussed extensively in our earlier work, we use the Fisher-Yates exact test, using as a measure of association strength either the p-value itself (cf. also Pedersen 1996), or the negative base-10 logarithm of the p-value (cf. Gries, Hampe & Schönefeld, 2005; Stefanowitsch & Gries 2005). Regardless of the distributional statistic used, the words $W_{1\text{-}n}$ occurring in (a given slot of) C can then be ranked according to the direction (attraction or repulsion) and the strength of their association to C. In previous work we have investigated the verb slots of, for example, the ditransitive construction (cf. (1)) and the *into*-causative (cf. (2)) to uncover systematic semantic patterns of attraction and repulsion (the relevant slot(s) are in bold):

(1) *I **gave** him a Jack Kerouac book.* (ICE-GB: S1A-015)

(2) *You wanted to **trick** us into believing that it was.* (BNC G0E)

We showed, for example, that earlier analyses of the ditransitive as a polysemous construction encoding ‘transfer’ and various related meanings are substantiated by collostructional analysis, or that the *into*-causative encodes ‘causation against the will of the causee’ (cf. Stefanowitsch and Gries 2003 for details).

The second method is *distinctive-collexeme analysis*: this method identifies those words that best distinguish between semantically or functionally near-equivalent constructions (for example, so-called syntactic alternations, cf. Gries & Stefanowitsch 2004a). The method is similar to the one described above, except that it compares the frequencies of $W_{1\text{-}n}$ in C not to their frequencies in the corpus as a whole, but to their frequencies in the corresponding slot in C’s near-equivalent (the method thus extends earlier work on distinctive collocates, cf. Church et al. 1991 and Gries 2003).

We have used this method to show, for example, that there are clear semantically motivated association patterns for the ditransitive vs. its prepositional dative counterpart, the *will*-future vs. the *going-to* future, and the two verb-particle constructions.

Most important for our present purposes, however, is the third method, covarying collexeme analysis (cf. Gries & Stefanowitsch 2004b, Stefanowitsch & Gries 2005). This method serves to investigate the interrelation between two different slots of the same construction and is thus somewhat similar to traditional collocation or colligation-based methods (except that, unlike the traditional methods, it pays close attention to syntactic and semantic structure). Instead of comparing the frequency of a word W in a construction C to its frequency in the corpus as a whole or in a near-equivalent of C, the method involves determining the frequency of a word in one slot of a given construction ($W_{slot\text{-}1}$) in relation to the words occurring in a different slot of the same construction ($W_{slot\text{-}2}$), as shown in Table 2. In other words, the method identifies $W_{slot\text{-}1}$/$W_{slot\text{-}2}$ pairs that cooccur significantly more frequently than expected.

	Word $W_{slot\text{-}1}$	Other Words in Slot 1	Row Totals
Word $W_{slot\text{-}2}$	**a**	b	**a+b**
Other Words in Slot 2	c	d	c+d
Column Totals	**a+c**	b+d	**N=a+b+c+d**

Table 2. Cooccurrence table for a covarying-collexeme analysis

We have applied this method to a range of constructions, including the verb and the gerund in the *into*-causative (cf. (3)), the two nominals in the English *s*-genitive (cf. (4)), and the verb and preposition in the *way*-construction (cf. (5)):

(3) *You wanted to trick us into believing that it was.* (cf. above)

(4) *Every person's situation is different.* (BNC A01)

(5) *The rest of us made our way to a farm.* (BNC G15)

Detailed discussions of our findings and their implications can be found in Gries & Stefanowitsch (2004b) and Stefanowitsch & Gries (2005); suffice it to say here that all three methods have proven valuable for the analysis of many different constructions and their semantics and also obtained a high degree of predictive power in the case of alternating pairs.

1.2 A Problem and a Potential Solution

In spite of the rewarding results and some experimental evidence in favor of collostructional analysis (cf. Gries, Hampe, & Schönefeld, this volume), there is a potential drawback in the way the method has been used so far. Most of the constructions we have investigated are frequent enough to occur with a vast number of different verbs with differing frequencies in the slots under investigation. Often (though not always) a large number of these verbs have statistically significant associations with the construction in question. For reasons of economy and exposition, we typically base our interpretation of the results on just the top twenty or top thirty significant collexemes, classifying them into semantically (more or less) coherent groups on the basis of intuition or common-sense criteria; this classification then plays a pivotal role in our discussion of the different subsenses of the construction in question, including the issue which of these is to be regarded as central. While we do not wish to deny the important role that interpretation and introspection will ultimately play in even the most rigorously empirical investigation, we would also not want to claim that our procedure is in any way ideal; obviously, a more bottom-up procedure would lend more objectivity, and thus credibility, to the empirical results.

The problem we face consists in determining the number and nature of semantic classes that are instantiated by a given set of words without recourse to prior assumptions or intuitions. While this problem has not received the attention it deserves—especially in the theoretically oriented linguistic literature—it is by no means new, and several treatments exist, especially in the computational-linguistic literature. Our problem is essentially that of the (inductive) identification of equivalence classes (also relevant in the identification of syntactic categories (e.g. Brill, Magerman, Marcus, & Santorini 1990), cooccurrence classes (e.g. Hindle 1990; Pereira, Tishby, & Lee 1993; Li & Abe 1996), sense classes (Gries, 2006; Divjak, 2006), or, as in our case, semantic classes (e.g. Waterman 1995; Schütze & Pedersen 1997; Schulte im Walde 2000; cf. also, of course, Levin 1993 as a theoretically-driven precursor to this idea).

A statistical technique that is often used to address such problems is that of (hierarchical) cluster analysis. Hierarchical clustering is a family of methods that aims at identifying and representing hierarchical (dis)similarity relations between n different items. Usually, this is done by (i) comparing pairwise (dis)similarities between the items in a (dis)similarity matrix, (ii) successively amalgamating all items into clusters such that the items within a cluster are highly dissimilar to each other and at the same time highly dissimilar from all other items and clusters, and (iii) representing the resulting structure in the form of a tree-like dendrogram (cf. Manning & Schütze 1999, Ch. 5 for an overview of uses of cluster analysis in computational linguistics). In order to apply a clustering approach in the context of collostructional analysis, we must, of course, decide on a basis for clustering. In computational linguistics—as in many traditional collocation-based studies—it is a common strategy to cluster node words on the basis of all their collocates within a user-defined span. However, many of these collocates will not be truly semantically related to the word(s) under investigation and will therefore contribute little but noise to the data points included into the analysis. The usual 'remedy' against this—again, as in much traditional collocation-based work—is to use a huge amount of data so that the theoretically interesting collocates will be frequent enough to allow for an identification of meaningful patterns. While this strategy will also be tested here (cf. Sections 2.1 and 2.2 below), it does not seem *a priori* promising in the context of collostructional analysis; constructions have to be identified (semi)-manually, so one cannot increase the size of a data set annotated for constructions as easily as one can increase that of a data set containing raw word forms which can easily be retrieved automatically. Instead, within the context of collostructional analysis a different approach is conceivable and much more promising: rather than including all available collocates of a particular collexeme, it seems plausible to include only its covarying collexemes (which, as discussed above, are words that occur in a well-defined slot of the same construction as the node word). This strategy is, in fact, a natural consequence of the high emphasis that collostructional analysis places on close attention to semantic and syntactic structure in general. It is also promising in light of our previous research on covarying collexemes: we have shown (Gries & Stefanowitsch 2004b, Stefanowitsch & Gries 2005) that the two verbs occurring in the *into*-causative (i.e. the finite main verb and the gerund embedded in the PP) or the verb and the preposition in the *way*-construction seem to interact systematically in various ways, which invites the inference that clustering the words in one slot according to the words in the other slot should yield more insightful results than an approach based on linear collocates.

2 Case Studies

In this section, we will investigate the ideas sketched out in the previous section in a series of three case studies. In Section 2.1, we will cluster the verbs occurring in the ditransitive construction on the basis of all their collocates within a span defined by sentence boundaries; in Section 2.2, we will apply the same strategy to the *into*-causative and compare it to the more precise strategy of clustering the same verbs on the basis of their covarying collexemes in the gerund slot; finally in Section 2.3, we will cluster the verbs occurring in the *way*-construction according to the prepositions they co-occur with (cf. examples (3-5) above). Let us briefly comment on the expected results. The main advantage of collostructional analysis over other approaches, demonstrated in previous work, is that it is more precise. One can therefore expect that, while clustering in general should already yield meaningful results, clustering on the basis of precisely-defined slots should increase the quality of the results dramatically (cf. in this context Levy, Bullinaria, & Patel's 1999 finding that even distinguishing between left and right collocates for clustering makes a substantial difference).

2.1 The Ditransitive

The first construction to be investigated here is probably the best known English argument structure construction, the ditransitive exemplified in (1) above. The most influential analysis of the ditransitive in a Construction Grammar framework is that of Goldberg (1995: Ch. 6), who analyzes the ditransitive as a polysemous argument structure construction with a central 'transfer' meaning and variety of sense extensions such as 'intended transfer', 'enabling of transfer', 'implied transfer', 'communication' (i.e., metaphorical transfer), etc. Such subsenses (or rather, the verb classes corresponding to them) should be identifiable if cluster analysis is indeed a feasible means to investigate constructional semantics.

We extracted all ditransitive clauses from the British component of the International Corpus of English (ICE-GB), a corpus of spoken and written British English of the 1990s that contains detailed and manually checked morpho-syntactic annotation (cf. http://www.ucl.ac.uk/english-usage/ice-gb/). All 1,824 verb tokens used ditransitively (including those with sentential objects) were identified (amounting to 87 verb types) as were all of their collocates, i.e., words occurring in the same sentence (amounting to 6,004 different collocate types). In order to reduce the sample to a manageable proportion of the corpus, and since it is well-known that clustering approaches yield their best results when applied to at least moderately frequent cases (cf. Levy & Bullinaria 2001 for evaluation), the analysis was restricted to the 32 most frequent verb types. For these 32 verb types and their 5,743 collocate types, a 5,743-by-32 cooccurrence table was constructed and submitted to a hierarchical agglomerative cluster analysis, where the similarity of the verb types in the columns was computed using the City-block (Manhattan) distance measure (cf. Levy et al. 1999 for justification) and clusters were amalgamated using Ward's method; for reasons of comparability the same parameters were of course used for all subsequent analysis. The resulting dendrogram is shown in Figure 1.

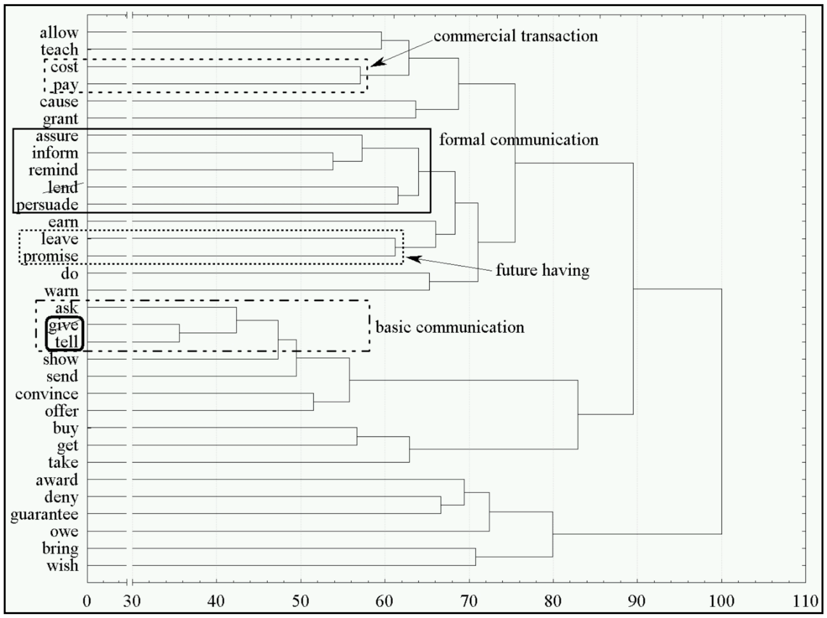

Figure 1. Dendrogram of the 32 most significant collexemes of the ditransitive (clustered according to sentential context)

As is obvious, there are several noteworthy clusters:

- the two commercial verbs *cost* and *pay* are grouped together;
- there is a cluster with several formal communication verbs: *assure*, *inform*, *remind*, *persuade*;
- the basic communication verbs *ask* and *tell* are grouped together;
- two verbs of future having are grouped together: *leave* and *promise*;
- verbs whose satisfaction conditions imply transfer are grouped together (*award*, *guarantee*, *owe*).[2]

In spite of these clusters, however, the overall results are not particularly encouraging. First, many clusters do not contain all verbs belonging to a corresponding semantic group; for example, the 'formal communication' cluster contains *persuade* and thus should also contain *convince*; but the latter is clustered together with the semantically unrelated *offer*. Second, two clusters contain verbs (formatted as strikethrough), which do not belong there semantically: *lend* is included in the 'formal communication' cluster and *give* is clustered together early with *ask* and *tell*

[2] We will comment on verbs framed with a rounded rectangle in Section 3 below.

(although the latter makes sense at least metaphorically). Finally, a considerable number of verbs are clustered in ways which are not semantically motivated or coherent at all, for example, *bring* and *wish*, *allow* and *teach*, and *cause* and *grant*.

It seems, therefore, that clustering on the basis of collocates alone is not particularly revealing especially if the data set is as small as the present one. Since the data set for the ditransitive cannot be increased easily, let us now turn to the results for the *into*-causative.

2.2 The *into*-causative

The *into*-causative as a partially lexically-filled construction was already exemplified above in (3); its syntactic and semantic form can be represented as [S_{Agent} $V_{\text{Causing-Event}}$ $O_{\text{Patient/Agent}}$ $_{\text{ADV}}$[*into* $\text{Gerund}_{\text{Resulting-Event}}$]] (Stefanowitsch and Gries 2003:224). In order to substantiate the semantic analysis (i.e., the classification of verbs) put forward in that study and to be able to validate the results on the basis of different data, 9,754 tokens of the *into*-causative from British and American were extracted from newspaper corpora.[3] These 9,754 tokens comprised 471 verb types and 26,579 collocate types, amounting to 5,300 different verb-collocate pairs. Again, the 32 most frequent verb types were extracted, which cooccurred together with different 21,276 collocate types (amounting to 66,012 cooccurrence tokens represented in a 21,276-by-32 cooccurrence table). This table was submitted to cluster analysis with the same parameters as before; the resulting dendrogram is represented in Figure 2.

[3] Two thirds of these data, the British data from the *Guardian*, were analyzed with a different focus in Gries & Stefanowitsch (2004b); for an analysis contrasting the British and the American data, cf. Wulff, Stefanowitsch, & Gries (2007); once again thanks are due to Britta Mondorf for providing the concordance.

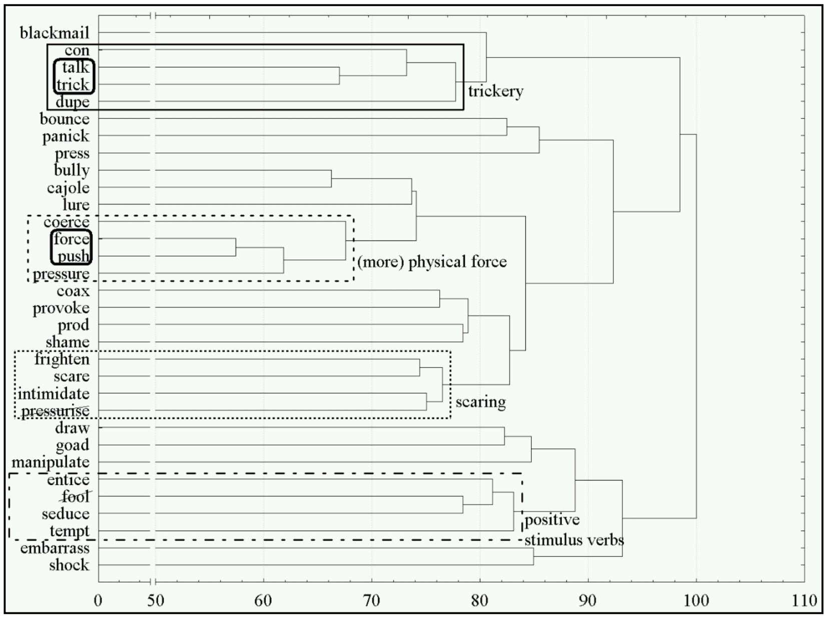

Figure 2. Dendrogram of the 32 most significant collexemes of the *into*-causative (clustered according to sentential context)

Again, a few clusters are worth mentioning with respect to their semantic patterning.

- there is a cluster with trickery verbs: *con*, *talk*, *trick*, *dupe*;
- there is a cluster with physical verbs of force: *coerce*, *force*, *push*, *pressure*;
- there is a cluster with some negative-persuasion verbs: *frighten*, *scare*, *intimidate*;
- there is a cluster with some positive-persuasion verbs: *entice*, *seduce*, *tempt*.

These clusters are rather more encouraging than those for the ditransitive, given that two of them are completely coherent (i.e., they do not have verbs in them belonging elsewhere); the results also tie in nicely with those of Stefanowitsch & Gries (2003, Section 3.2.1). However, there are, again, clusters which are incomplete and/or heterogeneous; for example, *press* and *bounce* should be included (or at least grouped more closely) with the 'physical force' cluster, *embarrass* and *shame* should be in the same cluster; *fool* does not fit with the 'positive stimulus' verbs, and some verbs are not part of any interpretable cluster at all (e.g. *bounce* and *panick*). In order to determine whether collexeme-based clustering is more successful, let us now consider the result of clustering the verbs in the *into*-causative in terms of their covarying collexeme, i.e. the gerund.

To that end, the same data set was analyzed into verbs and gerunds, yielding 471 verb types and 893 gerund types. Of these, again the 32 most frequent verbs were extracted, which cooccurred with 739 different gerund types (amounting to 6,800 cooccurrence tokens distributed in a 739-by-32 cooccurrence table), which was again submitted to a cluster analysis as described above. The dendrogram is shown in Figure 3.

Compared to Figure 2, these results are much more promising: There is just one major cluster of five verbs at the top which defies easy classification, but the vast majority of verbs is grouped into clusters which are very homogenous and do not contain verbs which—from the human analyst's semantic perspective—do not belong there:

- there are two clusters with physical force verbs: *coerce*, *force*, *push*, *pressure*, and *bounce* and *press*;
- there is the mini-cluster of provoking: *goad* and *provoke*;
- there is a cluster containing trickery (with the one exception of *talk*): *coax*, *trick*, *con*, *dupe*, *fool*;
- there is a cluster of positive-persuasion verbs: *entice*, *tempt*, *lure*, *seduce*;
- there is a cluster of six negative-persuasion verbs mostly involving fear: *embarrass*, *shame*, *panic*, *frighten*, *scare*, *intimidate*.

In sum, the cluster analysis yields a relatively coherent classification of verbs into semantic groups; moreover, these groups correspond fairly closely to the groups we posited in earlier work. While the covarying collexeme-clustering did not yield 'perfect' results, of course (note the clustering of *shock* and the 'positive stimulus' verbs), its superiority over the raw-collocate clustering is obvious.

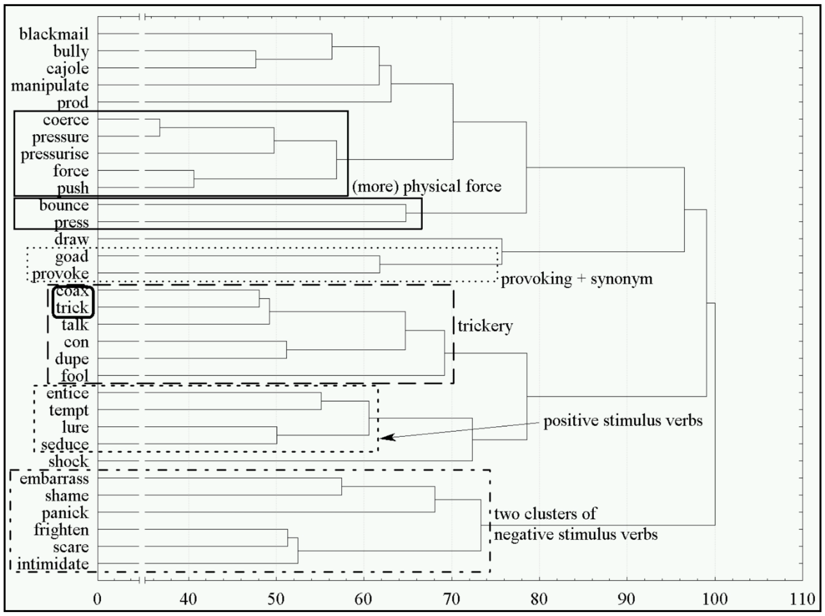

Figure 3. Dendrogram of the 32 most significant collexemes of the *into*-causative (clustered according to covarying verbs)

Note that in the case of the *into*-causative we have grouped together verbs on the basis of verbal collexemes; verbs are open-class items, and thus the number of collexeme types was very large. In the following section, we will attempt to cluster verbs in the *way*-construction according to prepositions; the latter form a closed-class, and the number of collexeme types will consequently be much smaller, which might affect the quality of the results.

2.3 The *way*-construction

The final construction to be investigated is the *way*-construction. It was exemplified in (5) above and can be represented as [S_{Agent} $V_{\text{create/move}}$ POSS $\textit{way}_{\text{path OBL}}$[P NP_{location}]]. It has also been analyzed in detail by Goldberg (1995, Ch. 9), who posited the disjunctive ‘create/move’ definition, and from a covarying-collexeme perspective by Stefanowitsch & Gries (2005).

For the analysis of the *way*-construction, we extracted all tokens from the British National Corpus 1.0; this yielded 5,831 tokens, comprising 492 different verb types and 214 different preposition types (including sequences of two or three prepositions, which were uniformly treated as complex lexical items, e.g. *up to* or *up on to*), yielding 1,569 verb-preposition pairs. In the previous case studies, the data entered into the hierarchical cluster analysis were restricted such that we clustered only those words which exceeded a particular frequency threshold on the basis of their covarying collexemes. An alternative suggestion by Levy & Bullinaria (2001) to increase the reliability of the analysis is not to delimit the number of words to be clustered, but the number of words on the basis of which the clustering is performed. In order to test to what degree the results of the cluster analysis can be further improved, we decided this time to combine both strategies and use the intersection of the 21 most frequent verb types and the most frequent 18 preposition types, which lead to an inclusion of 60% of all the data. This way, the proportion of cases included into the analysis is approximately the same as in the previous case studies, but the definition is slightly different. The resulting frequency table was submitted to the by now familiar cluster analysis, yielding the dendrogram represented in Figure 4.

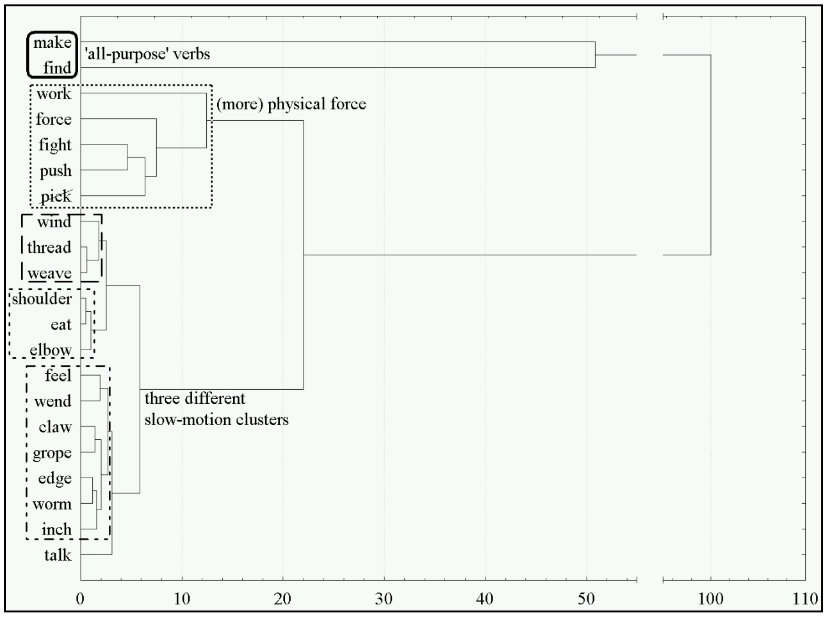

Figure 4. Dendrogram of the 21 most significant collexemes of the *way*-construction (clustered according to covarying prepositions)

Again the results are highly revealing. Nearly all of the verbs fall into a small number of relatively homogenous clusters:

- there is a cluster of the very general all-purpose verbs *make* and *find* (cf. below for discussion);
- there is a cluster of physical force verbs: *work*, *force*, *fight*, *push*;
- there is a super-cluster of slow movement verbs with three more specific smaller clusters: (i) a cluster with the non-linear movement verbs *wind*, *thread*, *weave*, (ii) a cluster with the body(-part) related motion verbs *shoulder*, *eat*, *elbow*, and (iii) a cluster with the careful/cumbersome movement verbs *feel*, *wend*, *claw*, *grope*, *edge*, *worm*, *inch*. The one unmotivated aspect of the entire dendrogram is the inclusion of *talk* in this cluster.

In other words, even though the data were selected differently and a closed-class item was chosen to cluster an open-class item, the clustering analysis yields very useful and clearly semantically motivated groups, providing us with valuable hints to potential subsenses of the construction.

3 Conclusions

On the basis of the results presented in the previous section, the prediction formulated above can now be evaluated. The evidence presented here clearly shows that clustering techniques can be a useful step towards making the semantic analysis of constructions more objective and more precise. This was to be expected given that clustering techniques have been applied to similar classification tasks successfully in the past. However, although clustering collexemes on the basis of unfiltered collocates may yield suggestive results, we have shown that the investigation of collexeme classes (and hence of constructional subsenses) by means of clustering really only becomes precise enough to be of interest for the theoretical linguist when it is combined with careful attention to semantic and syntactic structure. In other words, we have gone beyond the general expectation by showing that clustering based on frequent collexemes is far more precise and successful than previous approaches in which clustering is based only on a linear context or only on particular parts of speech within a particular window around the node word. In fact, the precision achieved by the manual identification of constructions and relevant slots within them seems to raise the quality of the results by such a substantial margin that it makes up for the relative sparseness of the data used here; in contrast to much work in computational linguistics, where clustering approaches often rely on many millions of data points, the present results were obtained on the basis of approximately 60% of the set of only a few thousand verbs and their collexemes.[4] Note in passing that the results support both frequency

[4] It must be borne in mind, of course, that clustering is a powerful exploratory technique which is sensitive even to minor differences in the data and the parameters chosen for the analysis. We have tried to select data and methods on the basis of proven and/or a priori justifiable

delimitations we used and that the quality of the clustering results was similar for a clustering based on open-class collexemes (Section 2.2) and one based on closed-class collexemes (Section 2.3).

Another issue is worth noting here, namely the high degree to which the present study supports previous collostructional analyses. Let us briefly return at this point to the verbs in the dendrograms above that were enclosed in a rounded rectangle; these were:

- *give*/*tell* for the ditransitive;
- *force*/*push* and *talk*/*trick* for the *into*-causative;
- *make*/*find* for the *way*-construction.

These verbs are all prominent in their trees, in that they are (i) merged earliest in the clusters they are part of and/or (ii) merged earliest (in the case of *give*/*tell*) or latest (in the case of *make*/*find*) in the whole of the tree. In other words, all these verbs somehow stand out; they occupy special positions within their trees. Interestingly, this cannot be explained with reference to their overall semantics in isolation since some of these verbs are relatively specific and some are very general; it can also not be attributed to their frequency within their respective constructions since they are not always the two most frequent ones in their construction. However, these verbs are exactly those which (i) instantiate the senses that can be considered the path-breaking verbs for their constructions (cf. Goldberg 1999) and (ii) have the highest collostruction strength for the major senses of their constructions. For example, the senses 'transfer' and 'communication' are arguably central to the ditransitive, and they are precisely the senses picked out by *give* and *tell*. Even more strikingly, recall Goldberg's analysis of the *way*-construction into two major senses, 'creation of a path' and 'movement along a path'; these senses are exactly the ones that are reflected by *make* and *find*. Finally, Stefanowitsch & Gries (2003, Section 3.2.1), provide evidence for the fact that, for the data in the BNC, causation by 'trickery' and by 'physical force' are the central senses of the *into*-causative, and the verbs *trick* and *force*/*push* correspond precisely to these senses. Interestingly, this is the case even though the data differs considerably in terms of variety (the present data contain one third of American English as opposed to the exclusively British English data in the BNC), medium (the present data set contains only written data while 10% of the BNC is spoken language) and register (the present data set exclusively contains journalese as opposed to the register-balanced BNC).

Note finally that the present findings also undermine all claims to the effect that the form-function relation of syntactic constructions and

criteria, but future research will show how successful we were in doing so.

argument structures (and the polysemy of constructions) cannot be learnt on the basis of the input to which the child is exposed during language acquisition. This is of course not meant to imply that the child performs cluster analyses, but if cluster analyses can yield such clear classes exclusively on the basis of input that is extremely sparse both quantitatively (i.e., in terms of the number of instances that enter into them) and qualitatively (since, unlike a child/learner who also gets additional semantic and contextual information, the cluster analyses uses no input other than collexeme frequencies), then this is prima facie evidence for assuming that the data exhibit enough hidden structure to allow access to constructional semantics (i.e. central senses and major extensions).

In sum, collostructional analysis has proven useful for a variety of reasons. It makes it possible to identify the verbs most strongly associated with constructions as well as their prominent sense extensions and verbs that are highly/most typical for constructions. Once the constructions and their collexemes have been identified, all these issues can be addressed objectively, rigorously, and easily.[5] Hopefully, this quantitative corpus-linguistic approach will be further applied to investigate constructions and their semantic as well as distributional characteristics.

References

Brill, E., D. Magerman, M. Marcus, & B. Santorini. 1990. Deducing Linguistic Structure from the Statistics of Large Corpora. *Proceedings of the DARPA Speech and Natural Language Workshop*, 275-282.

Church, K. W., W. Gale, P. Hanks, & D. Hindle. 1991. Using Statistics in Lexical Analysis. *Lexical Acquisition: Exploiting On-line Resources to Build up a Lexicon*, ed. U. Zernik, 115-164. Hillsdale, NJ: Lawrence Erlbaum.

Divjak, D. 2006. Ways of Intending: Delineating and Structuring Near-synonyms. *Corpora in Cognitive Linguistics: The Syntax-Lexis Interface*, eds. S. Th. Gries & A. Stefanowitsch, 19-56. Berlin: Mouton de Gruyter.

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. *The Emergence of Language*, ed. B. MacWhinney, 197-212. Mahwah, NJ: Lawrence Erlbaum.

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

[5] Software to perform collostructional analysis is available from the authors: PerlCLX 1.0 for collexeme analysis, distinctive-collexeme analysis and covarying collexeme analysis is available from Anatol Stefanowitsch; software to perform all these analyses as well as multiple distinctive collexeme analysis and clustering techniques of the kind presented here is available from Stefan Gries (Coll.analysis 3).

Gries, St. Th. & A. Stefanowitsch. 2004a. Extending Collostructional Analysis: A Corpus-based Perspective on 'Alternations'. *International Journal of Corpus Linguistics* 9: 97-129.

Gries, St. Th. & A. Stefanowitsch. 2004b. Covarying Collexemes in the *into*-Causative. *Language, Culture, and Mind*, eds. M. Achard & S. Kemmer, 225-236. Stanford, CA: CSLI.

Gries, St. Th. 2006. Corpus-based Methods and Cognitive Semantics: The Many Senses of *to run*. *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*, eds. St. Th. Gries & A. Stefanowitsch, 57-99. Berlin, Heidelberg, New York: Mouton de Gruyter.

Gries, St. Th., B. Hampe, & D. Schönefeld. 2005. Converging Evidence: Bringing Together Experimental and Corpus Data on the Association of Verbs and Constructions. *Cognitive Linguistics* 16: 635-676.

Gries, St. Th., B. Hampe, & D. Schönefeld. This volume. Converging Evidence II: More on the Association of Verbs and Constructions.

Hindle, D. 1990. Noun Classification from Predicate Argument Structures. *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 268-275.

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

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

Levy, J.P. & J.A. Bullinaria. 2001. Learning Lexical Properties from Word Usage Patterns: Which Context Words Should be Used? *Neural Network Models of Evolution, Learning and Development*, ed. R. French, 273-282. Berlin, Heidelberg, and New York: Springer.

Levy, J.P., J.A. Bullinaria, & M. Patel. 1999. Explorations in the Derivation of Cooccurrence Statistics. *South Pacific Journal of Psychology* 10: 99-111.

Li, H. & N. Abe. 1996. Learning Dependencies between Case Frame Slots. In *Proceedings of the 16th International Conference on Computational Linguistics*, 239-248.

Manning, C.D. & H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The M.I.T. Press.

Pedersen, T. 1996. Fishing for Exactness. *Proceedings of the SCSUG 96 in Austin, TX*, 188-200.

Pereira, F., N. Tishby, & L. Lee. 1993. Distributional Clustering of English Words. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 183-190.

Schütze, H. & J. O. Pedersen. 1997. A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval. *Information Processing and Management* 33: 307-318.

Schulte im Walde, S. 2000. Clustering Verbs Semantically According to their Alternation Behaviour. *Proceedings of the 18th International Conference on Computational Linguistics*, 747-753.

Stefanowitsch, A. & St. Th. Gries. 2003. Collostructions: Investigating the Interaction of Words and Constructions. *International Journal of Corpus Linguistics* 8: 209-243.

Stefanowitsch, A. & St. Th. Gries. 2005. Covarying Collexemes. *Corpus Linguistics and Linguistic Theory* 1: 1-43.

Stefanowitsch, A. & St. Th. Gries. 2009. Corpora and Grammar. *Handbook on Corpus Linguistics*, 933-951. Berlin/Heidelberg/New York: Mouton de Gruyter.

Waterman, S.A. 1995. Distinguished Usage. *Corpus Processing for Lexical Acquisition* eds. B. Boguraev & J. Pustejovsky, 143-172. Cambridge, MA: The MIT Press.

Wulff, S., A. Stefanowitsch, & St. Th. Gries. 2007. Brutal Brits and Persuasive Americans: Variety-specific Meaning Construction in the *into*-causative. *Constructing Meaning: From Concepts to Utterances*, eds. T. Berg et al., 265-281. Amsterdam & Philadelphia: John Benjamins.