

Quantitative approaches in usage-based cognitive semantics: myths, erroneous assumptions, and a proposal

Stefan Th. Gries
University of California, Santa Barbara

Dagmar S. Divjak
University of Sheffield

Abstract

In this paper, we assess objections formulated against (quantitative) corpus-linguistic methods in cognitive linguistics. We present claims critical of both corpus linguistics in general and particular corpus-linguistic analyses in particular and discuss a variety of theoretical as well as empirical shortcomings of these claims. In addition, we summarily discuss our recent corpus-based Behavioral Profile approach to cognitive semantics and illustrate its advantages in the domains of synonymy, polysemy, and cross-linguistic semantics as well as its methodological flexibility.

Key words

cognitive semantics, corpus linguistics, polysemy, synonymy, behavioral profile, English, Russian, statistical methods

1. Introduction

One of the most distinctive characteristics of cognitive linguistics is the prominent role that meaning and function play in linguistic analyses. This is true in two respects: first, in the sense that meaning and function are used to explain phenomena traditionally regarded as belonging to the domain of autonomous syntax; second, in the sense that studies of meaning and function themselves constitute a large body of cognitive linguistic work. At the very beginning, the methodology underlying cognitive-linguistic studies was quite homogeneous: just as in formal linguistics, analyses were nearly exclusively based on the acceptability or appropriateness of utterances that were often pulled out of their natural context and judged by the analyst him/herself. This admittedly rather unfortunate state of affairs began to change around the early 1990s when both experimental and observational approaches became more frequent. This change mirrored to some degree a general movement towards more rigorous empirical methods in linguistics, but was also facilitated from within cognitive linguistics itself by the recognition that the hallmark of cognitive linguistics – extremely fine-grained studies of the semantics of lexical items (in particular function words such as the classic studies of *over* and *there*) – suffered from a variety of methodological shortcomings (cf. Sandra and Rice 1995 for discussion). As a result, experimental research became more common in cognitive linguistics, in particular in idiom research by Gibbs and colleagues. It took another few years until observational approaches using corpus data took off. Their increasing popularity was no doubt facilitated by a rise of "usage-based approaches". As a result frequency-based accounts and corpus-linguistic methods gained ground, both in terms of publications, theme sessions at cognitive linguistics conferences, and in general visibility.

In spite of the omnipresence of advocates of usage-based approaches, corpus-linguistic approaches are still much less widespread in cognitive linguistics than one might be inclined to think. There are probably several reasons for this fact:

- new methodologies have never caught on fast in linguistics;
- corpus data are not always available and, if available, they tend to be difficult to handle, which is a friendly way of saying "introspective judgments of made-up data are so much easier to generate, keep track of, annotate, and evaluate" than thousands of diverse matches extracted from a corpus;
- corpus-linguistic assumptions and methods are often poorly understood;
- the quantitative underpinning of modern corpus-linguistic work does not come natural to linguists as the discipline, unlike psychology or psycholinguistics, has lacked a strong methodological foundation or awareness since the middle of the last century.

These issues conspired to yield yet another, attitudinal obstacle to a fast rise of corpus-linguistic methods: the conviction that corpus-linguistic methods have little to offer that the 'good ol' traditional approach' could not already do. We believe and will argue that this conviction, just like the issues that gave rise to it are entirely misguided. First, there is plenty of evidence suggesting introspective judgments come with a variety of difficulties that make them less objective, testable, replicable, and comparable with other studies than is desirable (cf. Labov 1972, 1975 as well as Schütze 1996 and the many studies cited therein). Second, the absence of quantitative underpinning makes it more difficult to compare such results with those of other studies, and the notion that some issues in linguistics are too complex and multidimensional in nature to allow analysis by mere introspection (cf. Gries 2003, Hinrichs and Szmrecsanyi 2007). Surely, psychologists and psycholinguists, who look at issues just as complex and noisy as linguists, jump through methodological and statistical hoops just for the fun of it.

In this largely programmatic and slightly polemic paper, we argue (i) against recent and less recent yet recurrent criticisms of corpus-linguistic methods and (ii) in favor of a particular corpus-based method for cognitive-semantic analyses, the so-called behavioral profile approach (extending an expression first used by Hanks 1996). More specifically, in the following section, we will first look at one attack against corpus-linguistic methods that we consider representative. Next, we briefly present our answer to this criticism, i.e. the behavioral profile method as well as outline and exemplify several of the advantages it has over competing approaches. Finally, we also address a variety of arguments often hurled at corpus-based approaches (within and outside cognitive linguistics) to bolster our point that corpus-based methods are probably the most useful yet most underestimated methodological tool available to contemporary (cognitive) linguistics.

2. Discussion

2.1 Criticism targeted at corpus-linguistic methods in cognitive linguistics

The most vociferous critic of corpus-based methods in cognitive linguistics we have come across is Raukko (1999, 2003). In two papers on English *get*, he argues vehemently against corpus-based methods in cognitive semantics (and in favor of his own experimental method). While a full-fledged rebuttal of the myriad of problems in his argumentation is not our concern (cf. Berez and Gries, ms, for that), his work exemplifies at least some of the above-mentioned problems.

One of these is Raukko's conception of corpus linguistics. This is how he characterizes the corpus-linguistic method:

The linguist looks at a large and somewhat pre-processed selection of text material and tries to find the relevant instances (instantiations, specimens) of the item that s/he wants to study. (Raukko 2003:165)

This statement is either a redundant truism, a severe misunderstanding or just as malign a misrepresentation. It is a redundant truism in the sense that, sure, if a corpus linguist investigates *get* in a corpus, he only looks for "relevant instances", i.e. instances of the verb (lemma) *get* and not for the noun *formaldehyde*. It is a severe misunderstanding or misrepresentation to think that a corpus linguist worthy of the name would look for instances of *get* in the corpus, yet would only classify those instances as relevant that do fit his theory instead of all of them or a representative randomized sample and thus avoid to deal with problematic instances and/or potential counterexamples. Put differently, contemporary corpus linguistics does not restrict itself to selecting those examples that fit a theory, disregarding the rest – on the contrary, it is a strength of the corpus-based approach, to which everybody who has ever looked at authentic data can testify, that a comprehensive corpus search typically results in data no introspection would have yielded and that all of these data are taken into account.

Raukko (1999:87) likewise takes issue with the fact that corpus linguists use introspection in their analysis of corpus data:

Other types of recent analyses of lexical polysemy [...] have made use of language corpora as sources of real-life data, but here also the analyst basically relies on her/his own linguistic introspection when analyzing the instances of a word in the texts and classifying them into neat semantic categories.

This statement is also either a redundant truism or a severe misunderstanding/misrepresentation. Of course, the analysis of corpus data requires classificatory decisions which are not always entirely objective – no corpus linguist in his right mind would deny this fact, just as no scientist would deny that some degree of intuition plays a role in nearly *any* study. The real issue is that corpus data often contain examples that an armchair linguist would not think of and, thus, force the researcher to take a broader range of facts into consideration. In addition, the concordance lines of a particular search expression and the uses of a word and their frequencies constitute an objective database of the kind that made-up sentences do not since researchers cannot invent all the uses of an expression in a corpus let alone their frequencies of occurrence. Thus, even if the classification of the data points is not always maximally objective, at least their nature, scope, and amount is, and the ideas underlying the annotation of examples can – and should – be made explicit. In addition, the corpus linguist will strive to analyze the entire set of to some extent subjectively annotated examples in an objective way (a point to which we will return later), postponing intuition until the stage of interpretation.¹

Finally, there are linguists who argue that introspection should be the central method of, say, cognitive semantics. In Talmy's (2000:4-5) words, "[c]ognitive semantics is thus a branch of phenomenology [...] the only instrumentality that can access the phenomenological content and structure of consciousness is that of introspection", but results from introspection "must be

correlated with those resulting from other methodologies" such as corpora, experimentation, and others. Again, we believe that subjective judgments are inevitable to some extent, yet the questions arise of when and how these judgments should be obtained and used. Talmy's (2000:6) argument that introspection "is already a necessary component in most of linguistics", e.g., in syntactic grammaticality judgments, is beside the point. The fact that many linguists have used introspection in the past does not mean there are no problems associated with researchers providing both theory and data, as we summarily discussed above. In both experimental and corpus-based studies, the primary source of data is not the analyst himself; it is a truism that data must still be interpreted, yet as many steps as necessary should be taken to avoid subjective biases, and theory-formation needs to be kept at least one step away from the retrieval of the data. We again ask: if nearly all cognitive psychologists and psycholinguists realize this, why is this so hard for many a linguist? True, corpus-linguistics studies meaning in terms of use, which in turn is made tangible through distribution, and hence lends itself better to quantification. Corpus-based approaches to meaning may not be able to capture the essence of abstract feelings like *love* or *faith*, but do other disciplines, typically considered to be better geared for this task, fare better in this respect? Has philosophy or religion come up with a generally accepted definition of either *love* or *faith*? And how well do these disciplines describe and predict when and how these concepts are used in everyday life? We strongly believe that cognitive linguistics can only benefit from reducing the subjective element in its methods as much as is feasible, and the methods and arguments presented below take important steps in this direction.

2.2 *The behavioral profile (BP) approach*

2.2.1 Introduction

As a corpus-based approach, the BP approach is based on the truism that corpus data provide (nothing but) distributional frequencies. A more relevant assumption, however, is that distributional similarity reflects, or is indicative of, functional similarity; our understanding of functional similarity is rather broad, i.e., encompassing any function of a particular expression, ranging from syntactic over semantic to discourse-pragmatic. The BP method involves the following four steps:

- the retrieval of (a representative random sample of) all instances of a word's lemma from a corpus in their context (usually at least the complete utterance/sentence);
- a (so far largely) semi-manual analysis of many properties of the use of the word forms; these properties are, following Atkins (1987), referred to as ID tags and comprise
 - morphological characteristics of the usage of the word in question: tense, aspect, mood, voice, number marking, etc.;
 - syntactic characteristics of the usage of the word in question: use in main or subordinate clauses, sentence type;
 - semantic characteristics: the sense of the word, semantic roles of the word's arguments and adjuncts;
- the generation of a co-occurrence table that specifies which ID tag level is attested how often in percent with each word (of a set of near synonyms or antonyms) or sense (of a polysemous word; the columns containing the percentages for each word or sense are then referred to as the word's or sense's behavioral profile. Consider Table 1 for an example.

ID tag		<i>begin</i>		<i>start</i>	
name	levels	<i>n</i>	%	<i>n</i>	%
sentence type	declarative	290	0.9732	511	0.9623
	interrogative	6	0.0201	12	0.0226
	imperative	2	0.0067	8	0.0151
clause type	main	135	0.453	231	0.435
	dependent	163	0.547	300	0.565
verb type	semi	128	0.4295	91	0.1714
	copula	0	0	1	0.0019
	transitive	0	0	2	0.0038
	monotransitive	34	0.1141	92	0.1733
	intransitive	118	0.396	243	0.4576
	semip	18	0.0604	102	0.1921

Table 1: An excerpt of the behavioral profiles for three ID tags of *begin* and *start*

- the evaluation of the table by means of descriptive techniques (such as summary frequencies), correlational methods, and exploratory cluster analysis.

In the following section, we discuss and exemplify several applications. The examples involve all kinds of statistical methods as well as examples from the domains of polysemy and near synonymy, within one language as well as across languages.

2.2.2 Applications in polysemy

Gries (2006a) uses descriptive methods based on behavioral profiles of the senses of *run* to address several of the central questions cognitive semanticists face. For example, Gries (2006a: Section 4.1) addresses the question of identifying the prototypical sense of *run* on the basis of several criteria including the most frequent sense and the formally least marked or constrained sense. Obviously, the BP approach allows operationalizing these two criteria straightforwardly. Classifying all concordance lines per verb sense makes it possible to count which sense is the most frequent one; the formally least constrained sense can be defined as the sense that is encountered with the largest attested number of ID tag levels (corrected for sense-frequency). Both criteria point to the sense 'fast pedestrian motion', which is not only intuitively correct, but also supported by other corpus-based, though not BP-based, findings such as the fact that this sense is both ontogenetically privileged (i.e., acquired first by children), phylogenetically privileged (i.e., one of two diachronically earliest senses) and in addition the most frequent sense of the zero-derived noun *run*.

BPs also answer the question of where to connect a particular sense of a polysemous word to the network of already identified senses. The example in question deals with the senses 'move away from something dangerous/unpleasant' and 'move away to engage in a romantic relationship',² the three most likely 'most likely' in the sense that they are semantically most similar – points of connection being the senses 'fast pedestrian motion', 'fast motion', and 'motion'. All other things being equal, Gries suggests to base one's decision of where to connect two senses on the overall distributional similarities between the two senses and between the candidate senses recognized in the network. The overall distributional similarity between two

senses is operationalized as the correlation coefficient of the two senses' behavioral profiles, and an investigation of all correlations between all senses shows that

- the five senses of *run* in question are much more similar to each other than all senses are to each other on average;
- the two senses that need to be connected are significantly more similar to 'fast pedestrian motion' than to 'fast motion' and 'motion', so this is how the network structure should be devised (in the absence of additional evidence to the contrary).

An example of a cluster-analytic approach in the domain of polysemy is Berez and Gries (ms). They investigate the senses of the highly polysemous verb *get* in a small sample of the the ICE-GB using the BP approach. They run a hierarchical agglomerative cluster analysis on the data and calculate *p*-values based on multiscale bootstrap resampling (cf. Shimodaira 2004, Suzuki and Shimodaira 2006). In spite of the small sample size, they find

- a cluster with all 'possess' senses ($p \approx 0.07$ ms);
- a cluster with all the 'acquire' senses ($p \approx 0.1$ ms);
- all non-causative 'move' senses ($p \approx 0.03$ *);
- a cluster that contains all causative senses (but also two other senses; $p \approx 0.21$ ns);
- a cluster that contains both grammaticalized senses 'must' and the *get*-passive (but also one other sense; $p \approx 0.08$ ms).

Four out of clusters are at least marginally significant, which is a good result given a small sample size and the fact that clustering is after all an exploratory method. The results therefore provide support for the fact that distributional characteristics are strongly correlated with semantic characteristics and senses of words, which in turn is exactly the assumption on which the whole BP approach is based.

2.2.3 Applications in near synonymy

In the domain of near synonymy, Divjak (2006) investigates five verbs that express 'intend' in Russian, whereas Divjak and Gries (2006) investigate nine Russian verbs meaning 'try' on the basis of the verbs' behavioral profiles. More specifically, the first study uses the BP approach to address the delineation (which verbs should be considered near-synonyms?) and structuring problem (how should a set of near synonymous words be structured?), whereas the second study focuses on the structuring and description problem (how can different words' meanings be compared reliably?).

Let us take a brief look at the try-study. On the basis of nearly 1,600 concordance lines of the verbs, their cluster-analytic approach reveals a tripartite cluster structure, but also reveals several interesting differences between the three clusters that are hard to discern in any other way. On the basis of *t*-values reflecting between-cluster differences, they show, for example, that the ID tag levels of each cluster give rise to a different abstract scenario (cf. Divjak and Gries 2006:42ff. and esp. Divjak, submitted, for details):

- the cluster $\{pytat'sja, starat'sja, probovat'\}$: a human is exhorted to undertaken an attempt to move himself or others (often negated);
- the cluster $\{silit'sja, proyvvat'sja, norovit'\}$: an inanimate subject undertakes several

- repeated but non-intense attempts to exercise physical motion;
- the cluster $\{pyzit'sja, tuzit'sja, tscit'sja\}$: an inanimate subject attempts in vain but intensely to perform what typically are metaphorical extensions of physical actions.

In addition, they use z -values to identify within-cluster differences because while, say, $\{silit'sja, proyvvat'sja, norovit'\}$ is a cluster distinct from the other verbs, that does of course by no means imply that the three are used identically.

An experimental follow-up study (Divjak and Gries submitted, for details see below) revealed that native speakers of Russian sort the nine verbs, with the exception of *silit'sja*, in exactly the same clusters as the corpus-based cluster analysis suggested. The fact alone that the BP approach is able structure the synonyms in a way that is so similar to the experimental results and that no intuitive lexicographic analysis has suggested so far shows how powerful corpus-based methods are and how useful the BP approach is in particular for cognitive-semantics.

Apart from these home-grown analyses that are BP analyses *per se*, there are some studies which differ from the BP approach quantitatively, yet not as much qualitatively. For example, Schmid's (1990) study of the phasal verbs *begin* and *start* involves a variety of ID tags similar to the ones mentioned above; the main difference to the BP approach is the degree to which statistical methods are used in the analysis.

Arppe's (2007) and Arppe and Järviö's (to appear) studies of Finnish verbs meaning 'think' is even more similar in that it involves both a very similar range of ID tags *and* multifactorial evaluation, but the foci are slightly different: in both papers, the focus is exclusively on the differences between words, i.e., a more coarse-grained approach than investigating both between-cluster and within-cluster differences, while the second paper is more concerned with showing how experimental data supplement corpus data (cf. below).

Finally, Dąbrowska (to appear) studies how nine English verbs of bipedal motion group together on the basis of their collocational patterns and semantic preferences. While she does not perform a detailed statistical analysis of the corpus results, our own cluster analysis of her data (with the same settings we used for all BP cluster analyses to date) is to some extent compatible with her own grouping.

2.2.4 Cross-linguistic studies

The studies in the previous section were all based on data from one language. Cross-linguistic semantic studies are notoriously difficult given that different languages carve up conceptual space(s) in different ways (cf. Janda, to appear, for discussion); for that reason linguistic dimensions are difficult to compare across languages. Since the BP approach is based on clearly operationalizable distributional properties, concordance lines from different languages can be annotated for a number of common characteristics while at the same time doing justice to any individual language's characteristics and avoiding overly subjective intuitions regarding cross-linguistic semantic differences.

Divjak and Gries (submitted a) study near synonymous phasal verbs in English (*begin* and *start*) and Russian (*natscat'*, *natscat'sja*, and *stat'*). As in other applications, they annotate concordance lines for these five verbs for a variety of criteria: morphological (tense, aspect, mode, person, voice), syntactic (clause, sentence, and complement types), argument-structural properties, semantic roles of subjects and complements as well as verb sense.

In a first part, Divjak and Gries investigate the synonyms within languages by comparing pairwise differences of the behavioral profiles. These comparisons show that the difference

between *begin* and *start* revolves around the semantic roles and characteristics of the Beginner and the Beginnee.

Interestingly, the main differences between the Russian verbs are not primarily concerned with the Beginner and the Beginnee. Rather, the verbs differ most strongly along aspectual and argument-structural lines. Thus, once one looks at the between-language differences, English and Russian phasal verbs opt for a different division of the conceptual space in question, and while such cross-linguistic distinctions may be overlooked in intuitive studies, they fall out immediately and naturally from their behavioral profiles.

Just as in the previous section, there is conceptually similar work, and Schönefeld (2006) is a case in point. She investigates translational equivalents of three basic posture verbs in English, German, and Russian. The main difference between her study and the BP approach outlined above is that she only includes collocations and not also morphosyntactic or semantic-role information. Another comparable study is Xiao and McEnery (2006), who explore near synonyms from three lexical fields (the consequence group, the cause group, and the price/cost group) on the basis of their collocational behavior in English and Mandarin Chinese.

2.2.5 Further methods, validation, and converging evidence

Since behavioral profiles are based on distributional properties captured by percentages, they offer possibilities that intuitive analyses usually lack: A final attractive feature of the BP approach, therefore, is the fact that it allows researchers to analyze the BP data using statistical techniques as well as to compare the results to data/results from other studies. Armchair data is much more limited in this respect.

For example, as we mentioned above, the BP approach was initially developed using descriptive, correlational, and exploratory, cluster-analytic quantitative methods. However, given the variety of studies we have mentioned, it is obvious that different techniques may well be used on the type of data collected in a BP. To name just two examples: To find out which variables drive the clustering Divjak (to appear) uses a linear discriminant analysis and in order to test the predictive power of the data contained in the behavioral profiles she fits a logistic regression model. Arppe (2007) and Arppe and Järvikivi (to appear) also use logistic regression to predict the choice of one near-synonym over another. To determine distinctive collocates, Schönefeld (2006) investigates her data using hierarchical configural frequency analysis (cf. von Eye 1990 and Gries 2004).

In addition, the quantitative nature of behavioral profiles allows for detailed comparisons of BP based results with experimental evidence. For example, Dąbrowska (to appear) is concerned with how data from a forced-choice selection tasks (of definitions and of video clips) and a gap-filling task relate to the collocational data discussed above. Arppe and Järvikivi (to appear) discuss their corpus data with results from a forced-choice selection task and an acceptability rating task.

More from the validation perspective, Divjak and Gries (submitted b) use a gap-filling task and a sentence-sorting task to test their BP-based cluster solution of the nine Russian verbs meaning 'try'. In the gap-filling task, subjects were given sentences from which the verb meaning 'try' had been deleted and which exhibited ID tag levels strongly associated with one verb. They then had to supply the verb they thought was most appropriate for the sentence. In the sentence sorting task, subjects were given sentences which differed only with regard to the verb and were asked to sort them into groups. Using chi-square tests and a similarity metric based on a Monte Carlo simulation, they found that the experimental findings are significantly more similar to the

BP-based cluster dendrogram than would be expected by chance, which lends strong support to the assumption that the BP approach yields cognitively realistic analyses.

By way of an interim summary, we have discussed applications and advantages of, as well as empirical evidence in favor of, the BP approach. Behavioral profiles can be used to investigate semantic relations of polysemy and synonymy at a high level of granularity and objectivity; they can be applied to simple cases with just two synonyms or larger sets with (so far) up to nine synonyms, where analysts' intuitions would become increasingly subjective, imprecise, and overtaxed; BPs allow to perform otherwise notoriously difficult cross-linguistic studies, and given their quantitative nature, they can be straightforwardly related to other empirical data and easily validated experimentally. Whichever limitations there may still be, we believe that the BP approach has much more to offer than many if not most other approaches to lexical semantics, and certainly more than some misguided and generic criticism of corpus-linguistic methods suggests.

2.3 *Criticism targeted at specific aspects of corpus-linguistic methods in cognitive linguistics*

So far, we have mainly been concerned with presenting advantages of the corpus-based BP approach in cognitive semantics (Sections 2.2.2 to 2.2.5) and disarming *general* points of critique raised against using corpora in cognitive semantics (Section 2.1). However, there is another set of arguments, leveled at *individual* corpus-based studies, both outside of and within cognitive linguistics. These can be summarized in what probably are the two most frequently heard and most disliked remarks after corpus-linguistic presentations:

- comments aimed at the corpus as a whole: "but isn't all this true in your corpus only?" or "I bet you would find something entirely different if you looked at a different corpus!" and "but the two corpora you are comparing are not sufficiently similar, your results are invalid!";
- comments aimed at subpart of the corpus: "I bet you would find something different if you looked at different registers!" or "I'm sure you would find something different if you looked at word forms/lemmas instead of lemmas/word forms."

In spite of their frequency, these comments tend to be invalid. First, they are theoretically problematic: The 'asker' hypothesizes a deviation from the null hypothesis (that there is no effect of or distributional difference between corpora), i.e., an alternative hypothesis, yet place the burden of proof on the 'askee'. If the asker thinks the distributional data obtained and reported on would be different in another corpus, the asker should test this alternative hypothesis instead of stipulating a difference for which (so far) no evidence exists; this is of course especially true when corpora on languages other than English are involved where alternative corpora are far from easily available (if at all).

Second, assertions like these are empirically problematic: The kinds of difference often hypothesized by askers is usually far from 'a given'. As a matter of fact, there now is an increasing amount of evidence that simple generalizations of what does and what does not remain constant across corpora, registers, word forms etc. are often inaccurate or exaggerated. Some of this evidence is based on BP type of approaches, other evidence is based on data regarding the distribution of occurrences of syntactic variables or the distribution of co-occurrences of lexico-syntactic variables.

As for the comments aiming at the corpus as a whole, for example, the results obtained

by Schmid (1993), who worked with the LOB corpus, are – while less comprehensive in terms of annotation and more comprehensive in terms of sense differentiation – to a considerably degree compatible with Divjak and Gries's (submitted a) results. This is noteworthy because the composition of the two corpora are of such a different nature that would compel many an audience to doubt the corpus comparability: Schmid's (1993) LOB consists exclusively of written and published texts representative for British English of the 1960s, whereas approximately 60% of the ICE-GB corpus used in Divjak and Gries (submitted a) consists of spoken language and even the 40% of written language in the ICE-GB contains a sizable amount of unpublished material.

Similar findings have been reported for the cherished distinction between spoken and written data. Stefanowitsch and Gries (to appear) and Gries (to appear) show that distinguishing between spoken and written data has no substantial effect in analyses of lexico-syntactic preferences of active vs. passive voice, the two word orders of verb-particle constructions, and the *will* vs. *going-to* future. Gries (to appear) shows that the same holds true for the ditransitive vs. prepositional dative alternation and that the 'real' division of the corpus – 'real' in the sense of explaining the maximally meaningful amount of variance in the corpus data as obtained by a principal component analysis – cuts across both spoken vs. written and all register distinctions present in the corpus. More specifically, the four corpus parts that are most homogeneous internally and most different from each other are based neither only on spoken vs. written nor only on subregisters; instead, they are mixed groups based on both these levels of granularity (cf. Gries, to appear, for details). This is of course something that linguists in general and linguists who have never adopted a bottom-up approach to corpus data in particular would be very reluctant to suggest; as scientists, they often prefer to stick to one level of categorization ... \$100 to the first corpus linguist who gets the following comment after a presentation: "Maybe you should forget about the mode and the registers – I bet the real distinctions are actually a mixture of different levels of corpus granularity." Gries (to appear) also finds that looking at word forms does not necessarily yield results different from a lemma-based analysis.

More generally, Gries (2006b) demonstrates on the basis of three very different case studies – the frequencies of the present perfect, the predictability of particle placement, and lexicosyntactic associations of the ditransitive constructions – that the usual suspects of mode, register and even subregister account for much less variability than the above-mentioned after-presentation comments suggest. In each of the above cases, different samples from even a single corpus may yield very different results; the size of within-corpus differences is often similar in size to between-corpus differences so there is little reason to *a priori* assume that other corpora will automatically yield different results. Bottom line: the issue of corpus homogeneity and comparability can only be determined (i) empirically and (ii) individually for each phenomenon, each corpus, and each level of corpus division(s) – it cannot be determined or objected *a priori* as one sees fit.

3. Concluding remarks

It goes without saying that this paper argues in favor of (in decreasing order of generality):

- multi-methodological approaches;
- corpus-linguistic approaches;

- BP approaches;

in linguistics in general and cognitive linguistics in particular. While *usage-based* is one of *the* buzzwords in contemporary cognitive linguistics, we believe that prototypical usage-based methods such as corpus-linguistic methods are still underutilized, misunderstood, misrepresented, and overcriticized, which is particularly interesting given that most analyses based on subjective and unfalsifiable intuitions by a native speaker-linguist are hardly ever subjected to any methodological critique ...

As far as multi-methodological approaches are concerned, it is probably fair to say that in linguistics as a whole the proportion of scholars using different methods is increasing at a steady pace; fortunately, this development has carried over to cognitive linguistics to some degree. The necessary development towards more quantitative methods, however, is progressing at a slower pace. The need of statistical methodological tools that are standard in most other social and cognitive sciences (!) has not yet been recognized uniformly.

That is to say, a usage-based linguistics needs quantification and statistical analysis. (Tummers, Heylen, and Geeraerts 2005:234)

The statistical analysis of empirical data, to be sure, should not be considered a fancy gadget designed to overwhelm linguists who are generally not really acquainted with statistical techniques. Instead, statistical techniques constitute an essential part of an empirical analysis based on corpus data. (Tummers, Heylen, and Geeraerts 2005:236)

For corpus-linguistic approaches, we have pointed out some points of critique fairly commonly leveled against both corpus-linguistic methods in general and corpus-linguistic studies in particular. We have addressed every single of these points theoretically and/or on the basis of empirical data and hope to have shown that these criticisms are often, though not always, false, biased, not substantiated, and premature. In addition, a variety of advantages of corpus-based approaches has hopefully become apparent: good corpus studies take into consideration all the variability that comes with many natural examples (as opposed to few judgments on potentially atypical examples), are gathered in an objective way and allow for replicability and validation (which intuitive judgments do not).

As for the BP approach, we have argued that this radically corpus-based approach, when applied to different kinds of semantic relations,

- yields more objective and more precise descriptive data than introspective analyses while at the same time staying true to the usage-based commitment of cognitive linguistics;
- allows for a bottom-up, data-driven study of distributional patterns on the basis of many quantitative techniques that outperform human analysts in terms of pattern recognition;
- allows for cross-linguistic comparisons of lexical semantics using objectively measurable distributional properties as opposed to difficult-to-port-across-languages semantic distinctions alone;
- allows integrating data and results from different sources and studies more easily than most other approaches (let alone intuitive approaches).

We invite the skeptic who thinks that the good ol' traditional way of doing semantic analyses can do all this and even more to illustrate how that is supposed to work ...

A final advantage of corpus-based approaches is that they are humbling; humbling in the positive sense that the sober reality of what is attested in corpora often puts a serious limit on bold theorizing. Put differently, the large degree of diversity corpus data exhibit as well as an indication of what is frequent and what is not, of what does and what does not reach or come close to reaching statistical significance, do not always support far-reaching theoretical models and force practitioners to take the usage-based perspective more seriously. Thus, given all the above we plead for not throwing out the corpus-linguistic baby with the argumentatively and methodologically muddy bathwater and hope that the incredibly powerful and flexible tool of quantitative corpus linguistics will become recognized for what it can and what it cannot do.

References

- Arppe, Antti. 2007. Multivariate methods in the corpus-based lexicography: a study of synonymy in Finnish. Paper presented at Corpus Linguistics 2007, University of Central England, Birmingham.
- Arppe, Antti and Juhani Järviö. 2007. Every method counts: combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3.2:131-59.
- Atkins, Beryl T. Sue. 1987. Semantic ID tags: Corpus evidence for dictionary senses. *Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary*, p. 17-36.
- Berez, Andrea L. and Stefan Th. Gries. ms. Experimental and corpus-linguistic evidence: compatible or competitors?
- Dąbrowska, Ewa. To appear. Words as constructions. In: Evans, Vyvyan and Stéphanie Pourcel (eds.). *New Directions in Cognitive Linguistics*. Amsterdam, Philadelphia: John Benjamins.
- Divjak, Dagmar S. 2006. Ways of intending: A corpus-based cognitive linguistic approach to near synonyms in Russian. In: Gries, Stefan Th. and Anatol Stefanowitsch. (eds.). *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*. Berlin, New York: Mouton de Gruyter, p. 19-56.
- Divjak, Dagmar S. submitted. *Structuring the lexicon: a clustered model for near-synonymy*. Berlin, New York: Mouton de Gruyter.
- Divjak, Dagmar S. and Stefan Th. Gries. 2006. Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2.1:23-60.
- Divjak, Dagmar S. and Stefan Th. Gries. submitted a. Corpus-based cognitive semantics: a contrastive study of phasal verbs in English and Russian. In: Lewandowska-Tomaszczyk, Barbara and Katarzyna Dziwirek (eds.). *Cognitive corpus linguistics studies*. Frankfurt am Main: Peter Lang.
- Divjak, Dagmar S. and Stefan Th. Gries. submitted b. Clusters in the mind? Converging evidence from near synonymy in Russian.
- von Eye, Alexander. 1990. *Introduction to Configural Frequency Analysis: the search for types and antitypes in cross-classifications*. Cambridge: Cambridge University Press.
- Gries, Stefan Th. 2003. *Multifactorial analysis in corpus linguistics: a study of Particle*

- Placement*. London, New York: Continuum Press.
- Gries, Stefan Th. 2004. HCFA 3.2. A program for R.
- Gries, Stefan Th. 2006a. Corpus-based methods and cognitive semantics: The many senses of *to run*. In: Gries, Stefan Th. and Anatol Stefanowitsch (eds). *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*. Berlin, New York: Mouton de Gruyter, p. 57-99.
- Gries, Stefan Th. 2006b. Exploring variability within and between corpora: some methodological considerations. *Corpora* 1.2:109-51.
- Gries, Stefan Th. to appear. Corpus data in usage-based linguistics: What's the right degree of granularity for the analysis of argument structure constructions?. In: Brda, Mario and Milena Žic Fuchs (eds.). *Expanding cognitive linguistic horizons*. Amsterdam, Philadelphia: John Benjamins.
- Gries, Stefan Th. and Dagmar S. Divjak. to appear. Behavioral profiles: a corpus-based approach to cognitive semantic analysis. In: Evans, Vyvyan and Stéphanie Pourcel (eds.). *New directions in cognitive linguistics*. Amsterdam, Philadelphia: John Benjamins.
- Hanks, Patrick. 1996. Contextual dependency and lexical sets. *International Journal of Corpus Linguistics* 1.1:75-98.
- Hinrichs, Lars and Benedikt Szmrecsanyi. 2007. Recent changes in the function and frequency of Standard English genitive constructions: a multivariate analysis of tagged corpora. *English Language and Linguistics* 11.3:437-74.
- Janda, Laura A. to appear. What is the role of semantic maps in cognitive linguistics? In: Piotr Stalmaszczyk and Wieslaw Oleksy (eds.). *Festschrift for Barbara Lewandowska-Tomaszczyk*.
- Labov, William. 1972. Some principles of linguistic methodology. *Language in Society* 1.1:97-120.
- Labov, William. 1975. Empirical foundations of linguistic theory. In: Austerlitz, Robert (ed.). *The scope of American linguistics*. Lisse: Peter de Ridder, p. 77-133.
- Miller, George. A. and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6.1:1-28.
- R Development Core Team. 2007. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <<http://www.R-project.org>>
- Schütze, Carson T. 1996. *The empirical base of linguistics: grammaticality judgments and linguistic methodology*. Chicago, IL: University of Chicago Press.
- Shimodaira, Hidetoshi. 2004. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Annals of Statistics*, 32:2616-41.
- Stefanowitsch, Anatol and Stefan Th. Gries. to appear. Register and constructional semantics: A collocation case study. In: Kristiansen, Gitte and René Dirven (eds.). *Cognitive sociolinguistics*. Berlin, New York: Mouton de Gruyter.
- Suzuki, Ryota and Hidetoshi Shimodaira. 2006. The pvclust package, version 1.2.0, URL <<http://www.is.titech.ac.jp/~shimo/prog/pvclust/>>.
- Talmy, Leonard. 2000. *Toward a Cognitive Semantics*. Vol. 1. Concept Structuring Systems. Vol. 2. Typology and Process in Concept Structuring. Cambridge, MA: MIT Press.
- Talmy, Leonard. 2007. Introspection as a methodology in linguistics. Plenary lecture at the 10th International Cognitive Linguistics Conference. Krakow, Poland.
- Tummers, Jose, Kris Heylen, and Dirk Geeraerts. 2005. Usage-based approaches in Cognitive Linguistics: a technical state of the art. *Corpus Linguistics and Linguistic Theory*

1.2:225-61.

Xiao, Richard and Tony McEnery. 2006. Collocation, semantic prosody, and near synonymy: a cross-linguistic perspective. *Applied Linguistics* 27.1:103-29.

- 1 Again, the fact that, in spite of Raukko's critique of corpus-based methods, his own approach fares no better in terms of objectivity will not be discussed here; cf. again Berez and Gries, ms.
- 2 Cf. Gries (2006: section 4.2) for why these are considered different senses in the first place.