# Behavioral profiles

## A fine-grained and quantitative approach in corpus-based lexical semantics

Stefan Th. Gries
University of California, Santa Barbara

This paper introduces a fairly recent corpus-based approach to lexical semantics, the Behavioral Profile (BP) approach. After a short review of traditional corpus-based work on lexical semantics and its shortcomings, I explain the logic and methodology of the BP approach and exemplify its application to different lexical relations (polysemy, synonymy, antonymy) in English and Russian with an eye to illustrating how the BP approach allows for the incorporation of different statistical techniques. Finally, I briefly discuss how first experimental approaches validate the BP method and outline its theoretical commitments and motivations.

In this paper, I will provide an overview of a recent approach towards corpus-based lexical semantics that tries to go beyond most previous corpus-based work, the so-called Behavioral Profile approach. The remainder of this first section provides a necessarily brief and general overview of previous traditional corpus-linguistic work in lexical semantics and mentions the shortcomings of such work and how the Behavioral Profile approach attempts to address them.

Lexical semantics is the domain of linguistics that has probably been studied most with corpora. The main assumption underlying nearly all corpus-based work in lexical (and constructional) semantics is that the distributional characteristics of a linguistic expression reveal many if not most of its semantic and functional properties. The maybe most widely-cited statement to this effect is Firth's (1957, p. 11) famous dictum that "[y]ou shall know a word by the company it keeps." However, other quotes may be actually even more explicit and instructive, such as Bolinger's (1968, p. 127) statement that "a difference in syntactic form always spells a difference in meaning" or Cruse's (1986, p. 1) statement that "the semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts." Most explicit in this regard is Harris (1970, p. 785f.):

> [i]f we consider words or morphemes *A* and *B* to be more different in meaning than *A* and *C*, then we will often find that the distributions of *A* and *B* are more different than the distributions of *A* and *C*. In other words, difference of meaning correlates with difference of distribution.

This kind of logic has been applied especially fruitfully in the domain of synonymy, where contextual information of two kinds has been particularly useful and revealing, co-occurrence information on the lexical level (i.e., collocations) and co-occurrence information on the lexico-syntactic and/or syntactic level (i.e., colligations). That is, on the one hand, synonyms were studied with regard to the different (sets of) words they co-occur with: cf. Kennedy (1991) on *between* versus *through*, Church, Gale, Hanks, and Hindle (1991) and Church, Gale, Hanks, Handle, and Moon (1994) on *strong* versus *powerful*, Partington (1998) on *absolutely* versus *completely* versus *entirely*, Biber, Conrad, and Reppen (1998) on *big* versus *large* versus *great*, Kjellmer (2003) on *almost* versus *nearly*, Taylor (2003) on *high* versus *tall*, Gries (2001, 2003c) on *alphabetic* and *alphabetical* and many other *-ic*/*-ical* adjective pairs; etc.; examples of work based on the same assumptions regarding distributions of collocates but on other lexical relations include Biber (1993) on polysemy (*right* and *certain*), Jones, Paradis, Murphy, and Willners (2007) for a set of antonyms; etc.

On the other hand, synonyms were also, but usually largely separately, studied in terms of their preferred grammatical associations: cf. Atkins and Levin (1995) on *quake* versus *quiver*, Biber et al. (1998) on *little* versus *small* or *begin* versus *start*, Gilquin (2003) on causative *get* and *have*, Wang (2006) on Mandarin *lian …* constructions, Arppe and Järvikivi (2007) and Arppe (2008) on several Finnish verbs meaning 'think'; examples of work based on the same assumptions regarding colligations but on other lexical relations include, e.g., Croft (1998, 2009) on the polysemy of *eat*.

Even though the above studies and many others have provided a wealth of evidence going beyond what 'armchair semantics' can provide, many of these studies still exhibit several areas of potential improvement. These can be grouped into three different categories: (i) the range of the elements that are studied; (ii) data and methods, and (iii) theoretical background. As for the first category, previous studies are sometimes limited such that nearly all corpus-based studies on synonymy or antonymy focus on only synonyms or antonyms and do not take larger sets of words, or words with very many different senses, into consideration. In addition, many studies focus only on the base forms of the words in question as opposed to including, or differentiating between, different inflectional forms of the relevant lemmas.

As for the second category, even though corpus data provide a wealth of distributional characteristics, many corpus-linguistic studies of lexical relations until relatively recently focus only on one of the two types of co-occurrence information: collocations *or* colligation, but adopt a very coarse-grained perspective both in terms of the number of distinctions made and in terms how little the two kinds of information are combined. For example, when it comes to collocations, previous work often either just includes all collocates in a user-defined window around the search word or collocates in a particular syntactically defined slot, and there are even studies that do not really make explicit which strategy was used (e.g., Taylor, 2003, who states that he included comparative and superlative forms of *high* and *tall*, but does not state how the collocates were identified).[1] This problem is exacerbated, in a sense, by the fact that most of the studies also do not analyze their distributional data in the most revealing way but rather restrict themselves to observed frequencies of co-occurrence, that is, they state how often which (kinds of) words or which syntactic patterns the synonyms or antonyms were observed with and infer from that some usually semantic characterization of how the words in question differ. Gries (2003c) is a case in point, but already somewhat more advanced because, unlike most other studies, he uses as a diagnostic statistic not just an observed frequency of some collocate, but a version of a *t*-value that has been tailored to identify distinctive collocates of, in this case, nouns immediately following say, *alphabetic* and *alphabetical*, *symmetric* and *symmetrical*, etc.

As for the final category, most of the above corpus-based studies remain descriptive and do not relate their findings to, or integrate them into, a more theoretical account by explaining what the findings 'mean' and what the theoretical or psycholinguistics commitments/presuppositions or implications of the findings would be.

The Behavioral Profile (BP) approach addresses the above three categories of problems. It is specifically geared towards the analysis of larger sets of synonymous/antonymous words, or highly polysemous words with many senses and in fact not only allows for, but specifically encourages, the inclusion of different forms of a lemma as well as very many different kinds of co-occurrence information (morphological, syntactic, semantic, functional, etc.). In addition, the BP approach integrates different kinds of statistical analysis, ranging from simple frequencies/percentages via correlations up to hierarchical cluster analyses and, by way of extension, logistic regression and is compatible with, and in part based on the logic of, exemplar-based approaches to language acquisition, representation, and processing. The following section outlines the BP methodology in more detail on the basis of recent work (cf. the section 'Behavioral Profiles: The Method') and discusses a variety of applications (cf. the remaining sections in 'The Method and Its Applications').

## The method and its applications

### *Behavioral profiles: The method*

The Behavioral Profile approach involves the following four steps:

Step 1: the retrieval of (a representative random sample of) all instances of the lemmas of the word(s) to be studied from a corpus in the form of a concordance; for Gries and Otani's (2010) study of English size adjectives, this concordance included the following examples from the British Component of the International Corpus of English (ICE-GB):

(1) a. I guess size is a *bigger* problem actually than funding (S1B-076)
 b. which have to be transmitted in the UHF portion of the spectrum because of the *large* amount of bandwidth required (W2B-034)
 c. […] our own *little* <,,> magic circle or whatever it is […] (S1A-027)

Step 2: a (so far largely) manual analysis and annotation of many properties of each match in the concordance of the lemmas; these properties are, following Atkins (1987), referred to as ID tags and include, but are not limited to, the morphological, syntactic, semantic, and other characteristics listed in Table 1:

**Table 1.** Examples of ID Tags and their Levels Used in Previous Studies

| Type of ID tag | ID tag | ID tag levels |
|---|---|---|
| morphological | tense | present, past, future |
| | mode | infinitive, indicative, subjunctive, imperative, and others |
| | aspect | imperfective, perfective |
| | voice | active, passive |
| | number | singular, plural |
| | person | first, second, third |
| | transitivity | intransitive, monotransitive, ditransitive, complex transitive, copular, … |
| | comparison | positive, comparative, superlative |
| | negation | affirmative, negative |
| syntactic | sentence type | declarative, exclamative, imperative, interrogative, … |
| | clause type | main, subordinate/dependent |
| | type/function of dependent clause | adverbial, appositive, relative, zero-relative, zero-subordinator, … |
| | modification | attributive, predicative |

**Table 1.** (*continued*)

| Type of ID tag | ID tag | ID tag levels |
|---|---|---|
| semantic | types of nominal arguments | abstract vs. concrete; or<br>animate (human, animal) vs. inanimate (concrete object, body part, event, phenomenon of nature, organization/institution, speech/text, …); or<br>count vs. mass |
| | types of verbal arguments | action, communication, emotions, intellectual activities, perception, …; or<br>accomplishment, achievement, process, state, semelfactive |
| | controllability of actions | high vs. medium vs. no controllability |
| | adverbial modification | no modification, locative, temporal, … |
| | sense | the sense of the polysemous word that is investigated |
| other | acceptability | yes, no |
| | collocates | the collocates of the word/sense that is investigated |
| | corpus | language1, language2, … (e.g., English vs. Russian); or<br>language1 as L1, language2 as L2 (e.g., native English vs learner English) |

A very small excerpt of the annotation resulting for the three examples listed in (1) is shown in Table 2.

**Table 2.** Examples of ID Tag Level Annotation

| Form | Syntax | Semantic type 1 | Semantic type 2 | Clause function | Clause level |
|---|---|---|---|---|---|
| *bigger* | attributive | count | abstract | OD | depend |
| *large* | attributive | non-count | quantity | NPPO | depend |
| *little* | attributive | count | organization/ institution | PU | main |

Step 3: the conversion of these data into a co-occurrence table that provides the relative frequency of co-occurrence of each lemma/sense with each ID tag level such that the percentages of ID tag levels sum up to 1 within each ID tag (cf. the rounded rectangles around cells in Table 3):

**Table 3.** Example of Behavioral Profile Vectors (for English Size Adjectives)

| ID tag | ID tag level | *big* | *great* | *large* | *bigger* |
|---|---|---|---|---|---|
| Syntax | adverbial | 0 | 0.01 | 0 | 0 |
| | attributive | 0.87 | 0.83 | 0.91 | 0.45 |
| | predicative | 0.13 | 0.16 | 0.09 | 0.55 |
| Modifiee_count | count | 0.94 | 0.71 | 0.98 | 0.95 |
| | non-count | 0.06 | 0.29 | 0.02 | 0.05 |
| … | … | … | … | … | … |

That is, each column represents a set of co-occurrence percentages for one word, or for one sense of a word, and this vector of co-occurrence percentages is called a Behavioral Profile (extending Hanks's [1996] term).

Step 4: the evaluation of the co-occurrence data of the type of Table 3 by means of statistical techniques such as pairwise difference of percentages, correlational approaches, and hierarchical cluster analysis. (A variety of aspects of steps 3 and 4 can be performed with the interactive R script BP 1.0, which is available from the author upon request.)

In the following sections, I will summarily discuss several examples of this approach that highlight its application to different lexical relations: examples involving polysemy are discussed in Sections 'The Polysemy of *to run*' and 'The polysemy of *to get*', applications in synonymy are addressed in Sections 'Russian Verbs Meaning 'to try'', 'Contrastive Phasal Verbs', and 'Size Adjectives', the latter section will also be concerned with the relation of antonymy. The data discussed in these sections come from English, Russian, and French and are analyzed using different statistical methods in step 4: co-occurrence percentages and correlations (Sections 'The Polysemy of *to run*' and 'Contrastive Phasal Verbs'), hierarchical cluster analysis (Sections 'The Polysemy of *to run*' and 'Size Adjectives') with multiscale resampling (Section 2.3), cluster analysis with *post hoc* evaluation of clusters (Section 'Russian Verbs Meaning 'to try''). In addition, I will discuss one example that is less typical of an application (since it does not involve step 3) but still a statistical evaluation based on the extremely fine-grained analysis of co-occurrence data using logistic regression (Section 'Case-by-case based Approaches to Alternations').

## *The polysemy of* ***To Run***

Some of the most difficult questions in the domain of polysemy involve the decisions of (i) whether to lump/split senses that appear both somewhat similar and somewhat different, and (ii) where to connect a sense to a network of already identified senses. In a study of the highly polysemous English verb *to run*, Gries (2006a) uses BP vectors to address these questions. 815 instances of the verb lemma *to run* from the Brown Corpus and ICE-GB were annotated with regard to 252 ID tag lemmas (including a large number of collocations), and the resulting table was transformed into a BP table of the kind exemplified in Table 3. However, deciding on which senses of *to run* to lump or split can be challenging. For example, the corpus data contained the three sentences in (2).

(2) a. and we ran back [$_{\text{GOAL}}$ to the car]
 b. Durkin and Calhoun came running [$_{\text{SOURCE}}$ from the post]
 c. I once ran [$_{\text{SOURCE}}$ from the Archive studio] [$_{\text{GOAL}}$ to the Start The Week studio]

While there is obviously a sense of *to run*, in fact its prototypical sense, that may be paraphrased as *fast pedestrian motion*, and while (2a) and (2b) do involve such *fast pedestrian motion*, these sentences differ regarding how the motion is profiled: the former elaborates on the goal of the motion and leaves the source unmentioned whereas the latter elaborates on the source, or origin, of the motion and leaves the goal unmentioned, which is information that the Behavioral Profile reflects. Following the general argument of distribution reflecting function and a more specific argument by Croft (1998), one can use that information to decide that the examples in (2a) and (2b) should not be considered separate senses: the corpus contains examples that contain both the goal argument of (2a) and the source argument of (2b), as exemplified in (2c). That is, semantically similar senses that share complementation patterns should be lumped because of that distributional similarity.

Consider, by contrast, (3) and (4).

(3) If Adelia had felt about someone as Henrietta felt about Charles, would she have run away [$_{\text{COMITATIVE}}$ with him]?

(4) He wanted to know if my father had beaten me or my mother had run away [$_{\text{SOURCE}}$ from home]

The sense in (3) can be paraphrased as *run away to engage in a romantic relationship* whereas the sense in (4) can be paraphrased as *run away from something unpleasant*. Again, the senses are similar but also different, but this time there is no attested example in the corpus data that combines a comitative and a source argument, which in turn suggests that the two senses should not be combined.

While these two examples involved checking only a small set of BP frequencies — those for a few complementation patterns — the logic of how similar senses are to each other can be extended to the whole vector. The data on *to run* contained several examples of an *escape* sense:

(5) a. When he loses his temper with her she runs off, taking young Jacob with her
b. The musician ran away from school when he was fifteen, but this escapade did not save him from the Gymnasium

One question is where, in a network of senses arising from a semantic analysis, this sense would be connected to: to *fast pedestrian motion* (because that is the prototypical sense and prototypical escapes involve such motion), or to the sense of *fast motion* or *motion* that are motivated by other senses in the network (because escapes need of course not involve pedestrian motion and because *motion*, for example, is a more general link)?

To answer such questions, one can compare the BP vectors for the relevant senses to each other using a straightforward correlation measure such as Pearson's *r*. In fact, if one computes all senses' intercorrelations, one obtains very straightforward results. The two senses in (6) are most dissimilar (yielding the lowest of all *r*s, 0.38), and the 'escape' senses of the type in (5) are in fact most similar to that of 'fast pedestrian motion', the overall prototype.

(6) a. their cups were already running over without us
b. he ran his eye along the roof copings

In sum, BP vectors can help determine whether senses should be lumped or split (by checking whether selected ID tag level combinations are attested or not, that is, have frequencies greater than 0) and where to connect senses to a network (by comparing senses' BP vectors using correlations).

### *The polysemy of 'to get'*

The question of similarities of senses can be also be approached from a broader perspective than just the pairwise comparison discussed in the previous section, and a particularly useful method in this connection is hierarchical cluster analysis. Hierarchical cluster analysis is an exploratory data analysis method that summarizes the similarities of data points to each other in a tree structure (dendrogram) that in turn is interpreted by a human analyst. Berez and Gries (2009) applied this method and a extension to be discussed below to another highly polysemous verb in English, *to get*. The annotated 600 instances of the lemma *to get* from the ICE-GB with regard to 54 ID tag levels. The 26 senses from the resulting co-occurrence table that were attested 5 or more times were then analyzed with a hierarchical cluster analysis (similarity metric: Canberra, amalgamation rule: Ward), which yielded a dendrogram of different groups of senses shown in the left panel of Figure 1. Several clusters emerge, including a cluster that captures most

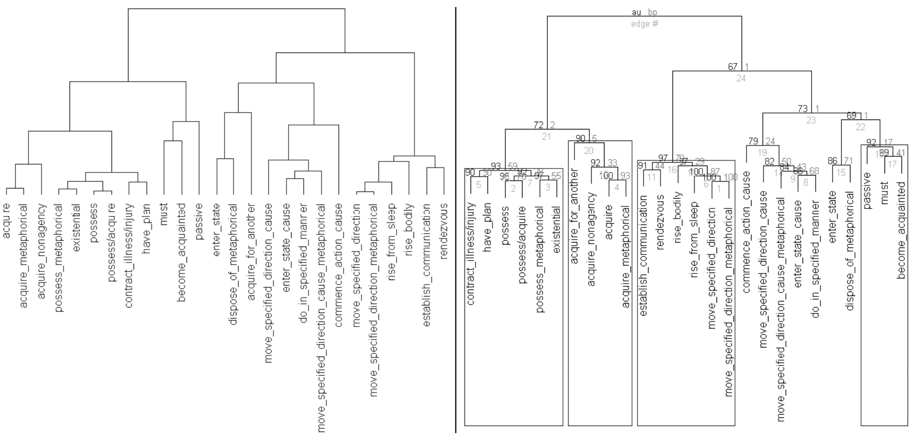


**Figure 1.** Dendrograms of 26 senses of to get in the ICE-GB: regular hierarchical cluster analysis (left panel); multiscale bootstrap resampling clustering (right panel)

*possession*-related senses, one with most *acquisition*-related senses, one with most 'movement' senses, and one with the more grammaticalized senses of *must* and the *get*-passive (Wright, Pollock, Bowe, & Chalkley, 2009, find very similar clusters in a validation study of *get* based on different corpus data.)

While dendrograms are sometimes very straightforward to interpret, it can sometimes be difficult to determine the number of clusters most strongly supported by the data. BP studies have therefore explored a variety of ways for follow-up analyses. The one to be mentioned in this section is a so-called multiscale bootstrap resampling (cf. Shimodaira, 2004, Suzuki & Shimodaira 2006), a method that applies multiple cluster analyses to resampled parts of the data and returns a dendrogram with *p*-values for all possible substructures. In the present case and in spite of the small sample, Berez and Gries (2009) obtained several suggestive clusters (a substructure with many movement senses and a substructure with all causative senses), but also some that were marginally or more significant:

- a 'possess' cluster, an 'acquire' cluster, and a cluster with the grammaticalized senses reach marginal significance ($p \approx 0.07$, $p \approx 0.1$, and $p \approx 0.08$);
- a non-causative 'move' cluster reaches significance ($p \approx 0.03$).

While such dendrograms will not always return all the clusters a semanticist might be interested in, this approach offers an objective way of narrowing down the space of possible ways in which senses are distributionally similar in authentic usage. The following section will discuss this kind of cluster-analytic approach in more detail, especially in terms of additional *post hoc* analysis.

*Russian verbs meaning 'to try'*

The previous section introduced hierarchical cluster analysis and its validation by resampling on the basis of a highly polysemous word. However, such clustering approaches can also be used for much smaller numbers of elements, such as when a several synonymous expressions are compared. Divjak and Gries (2006), for example, studied nine Russian verbs meaning *to try*. They annotated 1585 matches for 87 ID tag levels and converted the data into BP vectors that were then subjected to a hierarchical cluster analysis. The resulting dendrogram is represented in the left panel of Figure 2.

Again, the cluster analysis identifies a lot of structure in the data, but it is not completely clear whether the data consist of three or four clusters. In order to determine the most likely number of groups the synonymous verbs fall into, one can use the measure of average silhouette widths. Silhouette widths are a statistic that essentially compare within and between cluster similarities, and the higher an average silhouette width for a particular cluster solution, the better that cluster solution. Since nine elements can be grouped into between two and eight clusters, one way to determine the best possible cluster solution is to compute (average) silhouette widths for all possible numbers of clusters. The right panel of Figure 2 illustrates this approach: the silhouette widths on the *y*-axis are plotted against all possible numbers of clusters for nine elements on the *x*-axis (with black vertical lines and a grey step function for the averages). The graph shows that the cluster solution is best interpreted as exhibiting three clusters.

While the above approach supports the three-cluster interpretation of the above dendrogram assumed in Divjak and Gries (2006), further evaluation of the data is possible. To determine how the clusters differ from each other in terms of

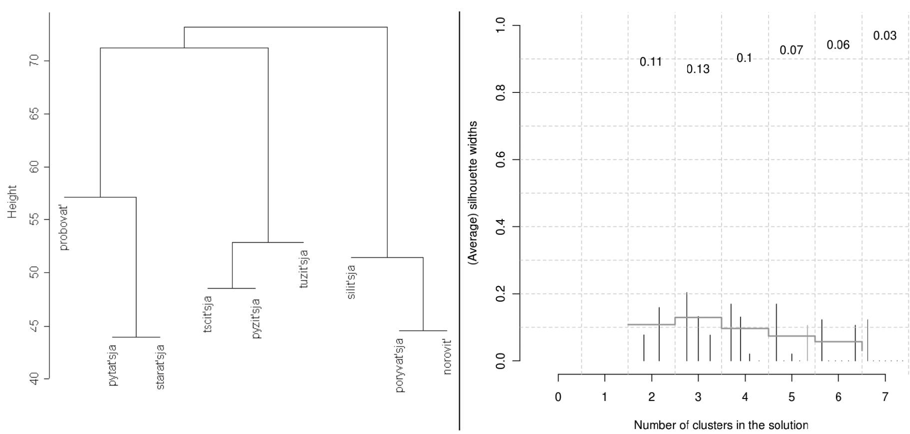


**Figure 2.** Dendrogram of nine Russian verbs meaning 'to try' (left panel) and a(verage) silhouette widths for all possible numbers of clusters (right panel)

the ID tags, they computed *t*-values that reflect which ID tags are overrepresented in one cluster compared to the others:

(7) $$t = \frac{mean_{\text{ID tag x of cluster c}} - mean_{\text{ID tag x}}}{standard\ deviation_{\text{ID tag x}}}$$

(cf. Backhaus, Erichson, Plinke, & Weiber, 2003, p. 534)

These *t*-values indicate clear differences between the three clusters that are compatible with some previous studies but at the same time more precise. The cluster {*poryvat'sja norovit' silit'sja*} is strongly associated with inanimate subjects, physical-motion verbs that often denote uncontrollable, repeated actions. By contrast, {*pyzit'sja tuzit'sja tschit'sja*} also features inanimate subjects, but more figurative physical-motion verbs affecting a second entity; crucially, this cluster is correlated with actions characterized by high vainness. Finally, {*probovat' pytat'sja starat'sja*} prefer animate subjects, but these are often exhorted to undertake attempt and, thus, perform it at reduced intensity. Crucially, these semantic characteristics are not based on an introspective analysis of individual contexts, but can largely be read off of the meanings of particles, adverbials, etc.

In sum, the *t*-values of the ID tags in the cluster solution allow for a straightforward identification of how (groups of) near synonyms differ entirely on the basis of their distributions in corpora. This type of cluster-analytic approach to BP vectors has now also been validated experimentally using both sorting and gap-filling experiments. Divjak and Gries (2008) show that, when Russian native speakers are asked to sort the nine Russian synonyms into groups, they exhibit a very strong and highly significant preference to sort them into groups that are compatible with the dendrogram in Figure 2. In addition, they also show that, when Russian native speakers are presented with sentences that exhibit ID tags with high *t*-values for a particular cluster but whose main verb meaning 'to try' has been deleted, then they are significantly more likely to fill that gap with a verb from the cluster for which the *t*-value 'primes' them. Both of these studies testify to the validity of (i) the overall approach, (ii) the procedure to identify the number of clusters using average silhouette widths, and (iii) the post hoc analysis of the clusters using *t*-values.

The next section will take this approach one level further. Since much of what is annotated are objectively countable linguistic features, both formal and functional in nature, this approach can actually be extended to cross-linguistic analysis, which will be discussed briefly in the next section.

*Contrastive phasal verbs*

Divjak and Gries (2009) use the BP approach to study near-synonymous phasal verbs in English and Russian. They retrieved 298 and 531 instances of *begin* and *start* from the ICE-GB respectively as well as 321, 173, and 156 instances of *načinat'/načat*, *načinat'sja/načat'sja*, and *stat'* from the Uppsala Corpus, annotated them for 73 ID tag levels, and created the usual type of co-occurrence table of BP percentages.

As a very simple way to explore the differences between the within-language synonyms, they computed the pairwise differences between BP percentages. For example, they computed the pairwise difference between the BP percentages of *begin* and *start* and identified the largest differences (which indicate ID tags much more frequent with *begin* than with *start*) and the smallest differences (which indicate ID tags much more frequent with *start* than with *begin*). They found that *begin* is more frequent than *start* in main clauses, with the progressive, and when nothing that is explicitly expressed or a concrete object initiates a change of state of itself or something abstract (events, processes, percepts). *Start*, on the other hand, is used more often than *begin* transitively, with *to*-infinitives, in subordinate clauses, and when a human instigator causes an action (particularly communicative actions) or, less so, causes a concrete object to operate. Similar computations and comparisons were performed for the three Russian verbs.

The more interesting aspect of this study, however, is that, to the extent that ID tags are applicable cross-linguistically, similar computations can be made to compare phasal verbs across languages. These provide two kinds of interesting information on how (dis)similarly words in different languages carve up semantic space: First, by comparing shared ID tag frequencies, one can see whether and/or to what degree a word in one language *corresponds to* another word in another language, and what the main differences between them are. Second, by comparing the kinds of ID tags that distinguish synonyms within one language to the kinds of ID tags that distinguish synonyms within the other language, one can see which parameters underlie lexical choices in different languages.

With regard to the former, Divjak and Gries (2009) find that English and Russian phasal verbs can only be mapped onto each other imperfectly. On the one hand, *begin* is similar to *načinat'/načat'* and *start* is somewhat similar to *stat'*: *begin* as well as *načinat'/načat'* prefer zero and more abstract beginners whereas *start/stat'* prefer past tense and similar beginnees (actions, communications, mental activities). On the other hand, both *begin* and *stat'* have features in common, too, as both highlight the view into the state after the onset of the action; the semantic features are differently grouped across verbs across languages.

With regard to the latter, the prototypes for each (set of) verb(s) revolve around different sets of characteristics. For example, 12 of the 15 most distinctive ID tags for *begin*/*start* involve beginners and beginnees, for example, *begin*'s preference for abstract processes and *start*'s concrete actions by humans. That is, the differences between *begin* and *start* are mainly lexico-semantic in nature. On the other hand, for the Russian verbs, lexico-semantic ID tags only constitute a minority among the most distinctive ID tags, which are much more concerned with aspectual and argument-structural properties of the verbs: For example, the Russian phasal verbs differ more with respect to the phase of action that is referred to and the agentivity.

In sum, the statistical method discussed here is rather simple *per se*, consisting, as it does, of mere differences between percentages. On the other hand, the annotation of cross-linguistically applicable ID tags allows for the precise study of subtle cross-linguistic formal and semantic/functional differences that are hard to identify in an armchair-semantic account or a study that involves a less fine-grained analytical approach.

### *Size adjectives*

One interesting test case is Gries and Otani's (2010) study of size adjectives in English. The study is interesting for two reasons, the first of which is concerned with the adjectives they study. They focus on a set of nearly synonymous size adjectives — *big*, *large*, and *great* — and a set of adjectives antonymous to those — *little*, *small*, and *tiny* — plus all their morphological forms attested in the ICE-GB. This setup is interesting because it included more than just two near synonyms but also because antonyms are potentially troublesome for distributional accounts: first, an antonym of a word *w* is the opposite of *w* so a theoretical approach based on the postulate that similarity of meaning is reflected in similarity of distribution may have to expect that *w* and its antonym would be distributionally very dissimilar. Second and on the other hand, an antonym of *w* is of course also somewhat similar to *w*, differing from it only on one semantic dimension, which would result in the opposite expectation of a high similarity of *w* and its antonym. Third, even if a word *w* is similar to its antonym, what if there are several antonyms and synonyms? For example, will *large* be more similar to *big* than to *great*? And will the BP approach confirm the finding that words have canonical antonyms and, for example, recognize that *large* is the canonical antonym of *small* and only a 'regular' antonym of *little*? Given such intricacies, data on sets of semantically related words like this provide a challenging test bed.

Gries and Otani (2010) retrieved and annotated 362, 409, and 609 instances of *big*, *large*, and *great* and their comparative/superlative forms respectively as well as

250, 409, and 34 forms of *little*, *small*, and *tiny* and their comparative/superlative forms from the ICE-GB. As usual, the concordance lines were annotated for ID tag levels (539), most of which were collocational, but they also included the usual morphological, syntactic, and semantic ID tags studied elsewhere. Then, the data were converted to BP vectors and submitted to the usual type of cluster analysis, the results of which are represented in Figure 3.

Obviously, the cluster analysis recovers a lot of very meaningful structure in the shape of both synonym-based clusters (e.g., {*smallest tiny*} and {*biggest largest bigger greatest*}) and perfectly canonical antonym-based clusters ({*big little*} {*large small*} and {*larger smaller*}), and these clusters are even morphologically very homogeneous. In addition, Gries and Otani computed the kind of pairwise differences between BP vectors mentioned above. The comparison between *big* and *large*, for example, showed that, compared to *large*, *big* prefers to modify non-count nouns, especially abstract nouns, but also sometimes humans and actions while *large* prefers count nouns, quantities but also organizations/institutions as well as non-human animate nouns.

Rather than discussing more specific results, it is worth emphasizing summarily here, that the BP approach alone yielded results from several previous studies, indicating how powerful a truly fine-grained corpus study can be:

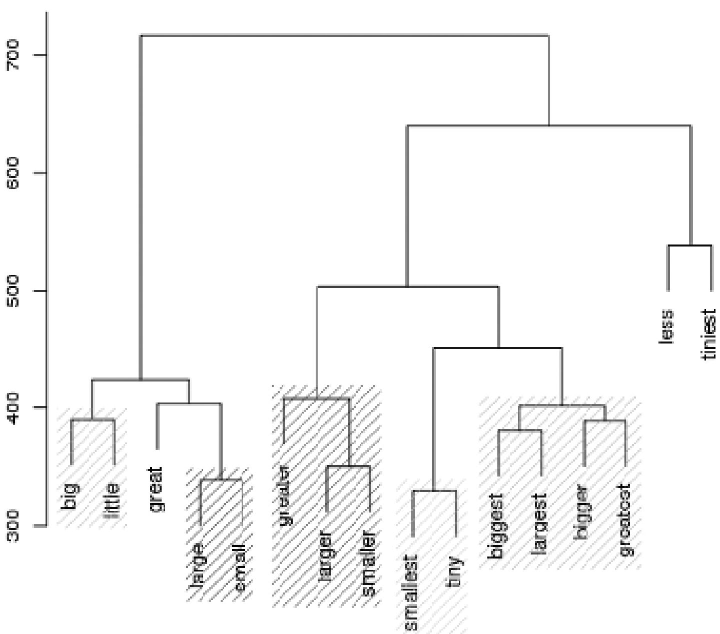


**Figure 3.** Dendrogram of behavioral profiles of selected size adjectives in the ICE-GB

- the close semantic similarity of *smallest* and *tiny* (as reported in Deese's 1964 study);
- the canonical antonym pairs (as reported in Jones et al. 2007, in whose corpus study *big* and *little* were not as strongly related);
- the morphologically clean clusters reflect subjects' preference to respond to a stimulus with a morphologically identical form (as reported in Ervin-Tripp, 1970).

The following section is concerned with the BP approach's theoretical commitments or implications and a comparison to an at least somewhat similar methodology.

## Behavioral profiles and their relation to other methods and theoretical accounts

The above applications discussed the Behavioral Profile approach proper and tried to highlight several of its advantages: It is based on authentic data, more specifically on a very fine-grained annotation of multiple linguistic dimensions, which is analyzed statistically. The results of BP approaches are often compatible, but also usually more precise, than previous work, and they have received first experimental support from three sorting and a gap-filling task. Given these results and advantages, this approach seems to work, but questions remain: why does it work and into what larger theoretical context can this approach be embedded? The answer is that the BP approach works because it taps into frequency information that is at the heart of contemporary exemplar-/usage-based models. The following section will outline some of the main assumptions of such models and why the Behavioral Profile approach yields the good results it does.

### *Exemplar-based models: Their main assumptions/characteristics and relation to BPs*

The main assumption of exemplar-based approaches is that each time a speaker processes a particular token/exemplar *E*, (aspects of) *E* is/are 'placed' in a hugely multidimensional space/network that comprises linguistic and encyclopedic knowledge. For example, phonemes are "associated with a distribution of memory traces in a parametric space, in this case a cognitive representation of the parametric phonetic space" (Pierrehumbert, 2003, p. 185). Such (distributional) characteristics of *E* involve

- phonetic, phonological, prosodic characteristics;
- morphological and syntactic characteristics;

- semantics and discourse-pragmatic characteristics;
- sociolinguistic characteristics;
- co-occurrence information of all aspects of *E*, involving both linguistic and extra-linguistic aspects (e.g., utterance contexts).

This multidimensionality raises the question of what learning, memory, and categorization look like in such an approach? Essentially, if a perceived *E* is close enough in multidimensional space to a cloud of already memorized exemplars (i.e., sufficiently similar to a category), then *E* will be 'added' into the multidimensional space at coordinates that represent its characteristics; in the case of vowel phonemes, for example, these could be formant frequencies etc. *E* will thereby strengthen the category formed by the already memorized exemplars to a degree proportional to its similarity to the cloud of already memorized exemplars and the homogeneity of the cloud of already memorized exemplars. Thus, "each instance redefines the system, however infinitesimally, maintaining its present state or shifting its probabilities in one direction or the other" (Halliday, 2005, p. 67). Consider, for instance, Figure 4, which shows the representations of traces of two linguistic elements in multidimensional space (reduced to two dimensions *x* and *y* for plotting).

Note that, first, the short black and grey rugs reflect the distributions of the ● and ▲ values on the *x*- and *y*-axes; second, horizontal and vertical lines indicate the means of the ● and ▲ values on the two plotted dimensions; and, third, the ● element is five times as frequent (with approximately the same dispersion as ▲) in memory. Consider now a speaker hearing a linguistic element that has the *x*- and y-coordinates of the point indicated by the grey/black X and the single longer rug line on each axis. Given the two dimensions of representation here, this element would probably be classified as an instance of ●: (i) X is closer to the means of ● than to the means of ▲ (as can be seen from the position of the X with regard to the solid lines representing the means) and (ii) ● is more frequent and more densely distributed around the coordinates of X than ▲ (as can be seen from the position of the single large rugs with regard to the other rugs). Once X has in fact been categorized as ●, then, graphically speaking, the X changes to a ●, hence updating the exemplar cloud, read 'memory representation', of ●.

This kind of approach has probably been fleshed out most in cognitive linguistics and psycholinguistics (cf. the work by Langacker, N. Ellis, Goldberg, Bybee, etc.) and explains findings from many different perspectives:

- developmental psycholinguistics, where acquisition follows from "exemplar learning and retention, out of which permanent abstract schemas gradually emerge […]. These schemas are graded in strength depending on the number of exemplars and the degree to which semantic similarity is reinforced

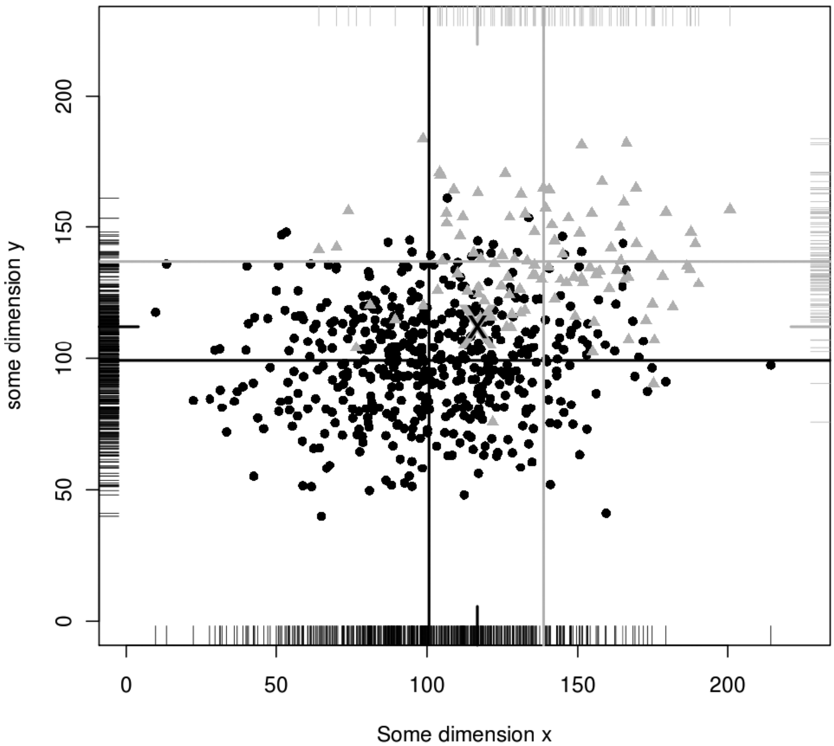


**Figure 4.** Schematic memory representations of traces of two linguistic elements ● and ▲ in multidimensional space (plotted on two dimensions *x* and *y*)

by phonological, lexical, and distributional similarity" (Abbot-Smith & Tomasello, 2006, p. 275);

- categorization/prototype effects, which follow from the multidimensional structure of an exemplar cloud: exemplars in the 'middle' of a cloud exhibit prototype effects;
- grammaticalization, where high-frequency tokens resist regularization due to their strong entrenchment; etc.

From the above characterization it should be very obvious that frequencies of occurrence and frequencies of co-occurrence are perhaps the single most important notion of such exemplar-based, or usage-based, approaches: memory traces of (co-)occurrences of events populate the multidimensional knowledge space and give rise to categorization and learning.

From this it should in turn also be obvious how much corpus-linguistic methods have to offer to this approach since, from some perspective at least, frequencies of occurrence and co-occurrence are all that corpora can ever provide. And while too much of corpus linguistics is still too atheoretical (cf. Gries, 2010 for discussion), some corpus linguists and psycholinguists have arrived at viewpoints that are highly compatible with the above approach. In a series of papers in the late 1980s/early 1990s, Miller and Charles developed the notion of a contextual representation, "a mental representation of the contexts in which the word occurs, a representation that includes all of the syntactic, semantic, pragmatic, and stylistic information required to use the word appropriately" (Miller & Charles, 1991, p. 26), which can be related to the above discussion straightforwardly. Hoey's (2005, p. 11) views are maybe even more explicitly related:

> the mind has a mental concordance of every word it has encountered, a concordance that has been richly glossed for social, physical, discoursal, generic and interpersonal context […] all kinds of patterns, including collocational patterns.

Thus, the BP approach works as well as it has so far because it is based on frequency information, which we have seen drives exemplar-based models, in particular the very fine-grained frequency information regarding the ID tags included in a specific BP study. In fact, this can be illustrated more clearly on the basis of Figure 4: the two dimensions *x* and *y* correspond to two ID tags (involving, in this case, continuous scales), and the BP frequencies of ID tag levels, which are used in correlations, pairwise differences, cluster analyses, etc. correspond to the *uni*dimensional distributions represented by the rugs on the axes.

Given this fit of the theoretical exemplar-based model and the corpus-based BP approach, it is not surprising that BP results provide new, interesting, and experimentally-validated findings that, especially for larger synonym/antonym sets and highly polysemous words, go beyond what many traditional corpus studies could reveal. There is one aspect, however, in which even the BP approach does not fully exhaust the multidimensionality of the data. Quantitatively-oriented studies of corpus data can informally be grouped as indicated in Table 4, which is sorted in ascending order of dimensionality.

In a nutshell, *n-dimensional* has two senses: *n*-dimensional$_1$ refers to the number of linguistic dimensions studied, and Table 4 distinguishes one, two, or multiple dimensions; *n*-dimensional$_2$ refers to the degree to which *co*-occurrence information is included, and Table 4 accordingly distinguishes occurrence-based approaches from co-occurrence-based approaches. In other words, a linguistic choice may be characterized with many ID tags, that is, involve many linguistic dimensions of variation, which would translate into an *n*-dimensional plot with *n* axes, and Figure 4 is a two-dimensional example in which the rugs reflect the distributions of the observed ID tag levels. That also means, as indicated in Table 4, the BP approach (in the highlighted row) is multidimensional$_1$: it takes into account that there are multiple levels of linguistic analysis on which expressions can be studied and that each of these may have a distinct frequency distribution. On

**Table 4.** Corpus-Based Studies on a 'Continuum of Dimensionality'[2]

| Variables/levels of analysis | Distributional approach |
|---|---|
| 1 | occurrence of variable levels: type/token frequencies (involving, e.g., chi-square tests of goodness-of-fit or similar tests) |
| 2 | occurrence of two variable levels: type/token frequencies (involving, e.g., chi-square tests of goodness-of-fit or similar tests) |
| 2 | co-occurrence of two variable levels (involving, e.g., chi-square tests for independence or correlations) |
| multiple (3+) | occurrence of multiple variable levels: summarized type/token frequencies (involving differences (Section 'Contrastive Phasal Verbs'), correlations (Section 'The Polysemy of *to run*'), clustering (Sections 'Russian Verbs Meaning 'to try'', 'Size Adjectives')) |
| multiple (3+) | co-occurrence of multiple variable levels: case-by-case observations (involving, e.g., regression approaches, correspondence analysis, hierarchical configural frequency analysis) |

the other hand, the BP approach is not multidimensional$_2$, because it does not preserve all of the co-occurrence information of all ID tags of each annotated data point on a case-by-case basis: the distribution of, say, the rugs on the two *x*-axes does not reveal what the *y*-axis values are for each *x*-axis value. For example, the BP approach for Figure 4 would reveal that there are many values for ID tag 1/dimension *x* that are close to 100, but it would not say what the corresponding *y*-values of these *x*-values are.

While this may seem as a potential downside of the BP approach, it is not necessarily one. First, the results discussed in Section 'The Method and Its Applications' and the list of advantages of the BP approach at the beginning of Section 'Exemplar-based Models: Their Main Assumptions/Characteristics and Relation to BPs' indicate that the BP approach is still vastly superior to many traditional descriptive corpus approaches in both its recognition of the need to incorporate more than just a few dimensions of variation, its quantitative rigor, and its close association to an increasingly supported theoretical approach. Second, there are situations in which the case-wise approach of regression studies is hard or even impossible to implement, namely when the data contain (i) (many) variables with many and/or infrequent levels and/or (ii) few data points. This is because trying to include *all* the co-occurrence information in such a table (recall Table 2 as an example) translates into including variables and all *n*-level interactions in the regression model, and in such designs, large numbers of unattested ID tag combinations can pose huge challenges to obtaining decent model fits and robust coefficient estimates. By contrast, a BP approach would remain largely unaffected by this since it only includes all *n* frequency distributions so that sparse co-occurrences will not arise. For example, in Divjak and Gries's (2006) study of nine Russian synonyms, two verbs were attested fewer than 100 times, which would result in very many unobserved higher-level interactions that would render a high-dimensional regression approach (esp. with many interactions) highly problematic.[3]

Nevertheless, since the case-by-case oriented approach is an important research strategy — particularly in the study of syntactic alternations — and can in fact complement a BP approach nicely, the following final section will briefly discuss this kind of design.

### *Case-by-case based approaches to alternations*

As argued above, the final application to be discussed here is somewhat different from all previous ones and the most specific implementation of the BP approach. All previous applications involved all four steps mentioned in Section 'Behavioral Profiles: The Method', thus including, most importantly, step 3, the computation of columns of relative frequencies of co-occurrence of a word/sense and some linguistic feature. The approach to be illustrated here does not involve step 3 but involves the application of some kind of regression-like statistical technique to the data table resulting from the annotation process of step 2. It is probably fair to say that this kind of approach has been used most productively in the study of syntactic alternations such as particle placement or the dative alternation, and that the first comprehensive study in this spirit is Gries's (2000) dissertation, published as Gries (2003a). Gries annotated a sample of verb-particle constructions from the British National Corpus — for example, *John picked up the book* versus *John picked the book up* — for a large number of what are here called ID tags and used a linear discriminant analysis (LDA) as well as classification and regression trees (CART) to determine which ID tags help predict native speaker constructional choices; an LDA was then also used in Gries (2003b) for the analysis of the dative alternation. While the use of an LDA is debatable — a binary logistic regression would probably have been better although the LDA results turned out to be not substantially different from the CART results and later logistic regression results on particle placement — there have now been several studies in the last few years that are based on a very similar logic, most famously perhaps, Bresnan, Cueni, Nikitina, and Baayen (2007). In addition, the range of applications has now begun to include not just syntactic alternations, but also semantic alternations.

For example, Arppe (2008) uses multinomial logistic regression (including a mixed effects model) to predict synonym choices, and Deshors and Gries (2010) as well as Deshors (in press) combine a BP approach with a binary logistic regression to study the use of *can* and *may* in English written by native speakers and by French learners (as well as *pouvoir* used by French native speakers). 3710 examples of these modals were retrieved from different corpora and annotated for 22 ID tags. They then first converted these data into BP vectors and studied the use of the modals with different cluster analyses, but they also applied a binary logistic regression to the *can* versus *may* data. More specifically, they tried to predict the choice of *can* vs. *may* based on all annotated ID tags and, crucially, the ID tags' interaction with a variable Corpus (with the two levels *NativeEnglish* and *LearnerEnglish*). The logic behind this approach this is that if a particular ID tag participates in a significant interaction with the variable Corpus, then this means that the ID tag's effect on the choice of *can* vs. *may* differs between native speakers and learners. Indeed, Deshors and Gries obtained not only a very high overall correlation ($R^2 = 0.955$; $p < 0.001$) but also found that several ID tags interact with Corpus in a way that is compatible with processing principles such as Rohdenburg's complexity principle.

In sum, there are occasions where the case-by-case approach using fine-grained annotation can provide results that very nicely complement the BP approach to the same annotation discussed here even within one and the same study. It is again important to note, however, that even these examples of regression approaches also had to use only a very small degree of dimensionality/interactions to avoid the data sparsity issues mentioned above. Thus, while the BP approach does not provide the most high-dimensional resolution possible, it avoids that particular problem, and I hope to have shown that it has a variety of very attractive features: it achieves a high degree of descriptive power, has received experimental support, is fully compatible with the widely supported theoretical and explanatory approach of exemplar-based models, and can nicely complement other statistical approaches in corpus-based semantics or computational linguistics.

## Notes

**1.** Some early studies that use more than just collocational or just syntactic information are cognitive-linguistic in nature. For instance, Schmid (1993) studied many lexical and syntactic characteristics of *begin* and *start* in an exemplary fashion. Also, in cognitive linguistics, Kishner and Gibbs (1996) studied collocations and syntactic patterns of *just*, and Gibbs and Matlock (2001) investigated uses of the verb *make*. Corpus-linguistic studies that are also appreciably broader in scope are Atkins (1987) study of *risk*, Hanks's (1996) study of *urge*, and Arppe and Järvikivi (2007), who all involved collocate and/or colligation analysis at an otherwise rare level of detail.

**2.** In all approaches on this continuum, type/token frequencies can, in fact *should*, also be studied with an eye to the dispersion of the variable levels in the corpus or its parts. However, in spite of the indisputable relevance of the notion of dispersion, this issue is unfortunately still somewhat understudied (cf. Gries, 2006b, 2008, 2009 for discussion).

3. Advocates of regression approaches may claim that regressions may become more feasible if levels of variables with many levels are conflated. While that is correct and may, if applied relentlessly, allow a regression approach, such conflations would of course also result in a loss of information and it is not clear at all whether the inclusion of hard-to-interpret higher-level interactions outweighs the loss of precision at the level of the individual variables.

## References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review, 23*, 275–290.

Arppe, A. (2008). Univariate, bivariate and multivariate methods in corpus-based lexicography — a study of synonymy. Unpublished Ph.D. Dissertation, University of Helsinki, URN: <http://urn.fi/URN:ISBN:978-952-10-5175-3>.

Arrpe, A, & Järvikivi, J. (2007). Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory, 3*, 131–159.

Atkins, B. T. S. (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*, 17–36.

Atkins, B. T. S., & Levin, B. (1995). Building on a corpus: a linguistic and lexicographical look at some near-synonyms. *International Journal of Lexicography, 8*, 85–114.

Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2003). *Multivariate Analysemethoden: eine anwendungsorientierte Einführung* [Multivariate analytical statistics: a practical introduction]. 10th edition. Berlin and New York: Springer.

Berez, A. L., & Gries, St.Th. (2009). In defense of corpus-based methods: a behavioral profile analysis of polysemous *get* in English. In S. Moran, D. S. Tanner, & M. Scanlon (Eds.), *Proceedings of the 24th Northwest Linguistics Conference. University of Washington Working Papers in Linguistics Vol. 27* (pp. 157–166). Seattle, WA: Department of Linguistics.

Biber, D. (1993). Co-occurrence patterns among collocations: A tool for corpus-based lexical knowledge acquisition. *Computational Linguistics, 19*, 549–556.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.

Bolinger, D. L. (1968). Entailment and the meaning of structures. *Glossa, 2*, 119–127.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G. Boume, I. Kraemer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 69–94). Amsterdam: Royal Netherlands Academy of Science.

Church, K. W., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisition: exploiting on-line resources to build a lexicon* (pp. 115–164). Hillsdale, NJ: Lawrence Erlbaum.

Church, K. W., Gale, W., Hanks, P., Hindle, D., & Moon, R. (1994). Lexical substitutability. In B. T. S. Atkins & A. Zampolli (Eds.), *Computational approaches to the lexicon* (pp. 153–177). Oxford and New York: Oxford University Press.

Croft, W. (1998). Linguistic evidence and mental representations. *Cognitive Linguistics, 9*, 151–173.

Croft, W. (2009). Connecting frames and constructions: a case study of *eat* and *feed*. *Constructions and Frames, 1*, 7–28.

Cruse, D. A. (1986). *Semantics*. Cambridge: Cambridge University Press.

Deshors, S. C. (2011). A multifactorial study of the uses of *may* and *can* in French–English interlanguage. Unpublished Ph.D. dissertation, University of Sussex.

Deshors, S. C., & Gries, St. Th. (to appear). A case for the multifactorial assessment of learner language: the uses of *may* and *can* in French-English interlanguage. In D. Glynn & Justyna Robinson (Eds.), *Polysemy and synonymy: corpus methods and applications in Cognitive Linguistics*. Amsterdam and Philadelphia: John Benjamins.

Deese, J. (1964). The associative structure of some common English adjectives. *Journal of Verbal Learning and Verbal Behavior, 3*, 347–357.

Divjak, D .S., & Gries, St. Th. (2006). Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory, 2*, 23–60.

Divjak, D. S. & Gries, St. Th. (2008). Clusters in the mind? Converging evidence from near synonymy in Russian. *The Mental Lexicon, 3*, 188–213.

Divjak, D. S., & Gries, St. Th. (2009). Corpus-based cognitive semantics: A contrastive study of phasal verbs in English and Russian. In K. Dziwirek & B. Lewandowska-Tomaszczyk (Eds.), *Studies in cognitive corpus linguistics* (pp. 273–296). Frankfurt am Main: Peter Lang.

Ervin-Tripp, S. (1970). Substitution, context, and association. In L. Postman & G. Keppel (Eds.), *Norms of word association* (pp. 383–467). New York: Academic Press.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. Reprinted in F. R. Palmer (Ed.). (1968). *Selected Papers of J. R. Firth 1952–59*. London: Longmans.

Gibbs, R. W. Jr., & Matlock, T. (2001). Psycholinguistic perspectives on polysemy. In H. Cuyckens & B. Zawada (Eds.), *Polysemy in Cognitive Linguistics* (pp. 213–239). Amsterdam and Philadelphia: John Benjamins.

Gilquin G. (2003). Causative 'get' and 'have'. So close, so different. *Journal of English Linguistics, 31*, 125–148.

Gries, St. Th. (2000). *Towards multifactorial analyses of syntactic variation: the case of Particle Placement*. Ph.D. dissertation, University of Hamburg.

Gries, St. Th. (2001). A corpus-linguistic analysis of *-ic* and *-ical* adjectives. *ICAME Journal*, *25*, 65–108.

Gries, St. Th. (2003a). *Multifactorial analysis in corpus linguistics: a study of Particle Placement*. London and New York: Continuum Press.

Gries, St. Th. (2003b). Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics, 1*, 1–27.

Gries, St. Th. (2006a). Corpus-based methods and cognitive semantics: the many meanings of *to run*. In St. Th. Gries & A. Stefanowitsch (Eds.), *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis* (pp. 57–99). Berlin and New York: Mouton de Gruyter.

Gries, St. Th. (2006b). Exploring variability within and between corpora: some methodological considerations. *Corpora, 1*, 109–151.

Gries, St. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics, 13*, 403–437.

Gries, St. Th. (2009). Dispersions and adjusted frequencies in corpora: further explorations. In St. Th. Gries, S. Wulff, & M. Davies (Eds.), *Corpus linguistic applications: current studies, new directions* (pp. 197–212). Amsterdam: Rodopi.

Gries, St. Th. (2010). Corpus linguistics and theoretical linguistics: a love-hate relationship? Not necessarily … *International Journal of Corpus Linguistics, 15*, 321–337.

Gries, St. Th., & Otani, N. (2010). Behavioral profiles: a corpus-based perspective on synonymy and antonymy. *ICAME Journal, 34*, 121–150.

Halliday, M. A. K. (2005). *Computational and quantitative studies*. London, New York: Continuum.
Hanks, P. (1996). Contextual dependency and lexical sets. *International Journal of Corpus Linguistics, 1*, 75–98.
Harris, Z. S. (1970). *Papers in structural and transformational linguistics*. Dordrecht: Reidel.
Hoey, M. (2005). *Lexical priming: a new theory of words and language*. London and New York: Routledge.
Jones, S., Paradis, C., Murphy, M. L., & Willners, C. (2007). Googling for 'opposites': a web-based study of antonym canonicity. *Corpora, 2*, 129–154.
Kennedy, G. (1991). *Between* and *through*: the company they keep and the functions they serve'. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics* (pp. 95–110). London: Longman.
Kishner, J. M., & Gibbs, R. W. Jr. (1996). How *just* gets its meanings: polysemy and context in psychological semantics. *Language & Speech, 39*, 19–36.
Kjellmer, G. (2003). Synonymy and corpus work: on *almost* and *nearly*. *ICAME Journal, 27*, 19–27.
Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes, 6*, 1–28.
Partington, A. (1998). *Patterns and meanings: using corpora for English language research and teaching*. Amsterdam and Philadelphia: John Benjamins.
Pierrehumbert, J. (2003). Probabilistic phonology: discrimination and robustness. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 177–228). Cambridge, MA: The M.I.T. Press.
Schmid, H.-J. (1993). Cottage *and* co., idea, start *vs*. begin. Tübingen: Max Niemeyer.
Shimodeira, H. (2004). Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Annals of Statistics, 32*, 2616–2641.
Suzuki, R. & Shimodeira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics, 22*(12), 1540–1542.
Taylor, J. R. (2003). Near synonyms as co-extensive categories: *high* and *tall* revisited. *Language Sciences, 25*, 263–284.
Wang, C. (2006). What distinguish synonymous constructions: A corpus-based study of Mandarin *lian … dou* and *lian … ye* constructions. Paper presented at GCLA 2006.
Wrigt, D., Pollock, L., Bowe, L., & Chalkley, W. (2009). Corpus-based investigation of polysemy: validating and extending the Behavioral Profile approach: a case study of the verb *to get*. Poster presented at WATM 2009.