Behavioral profiles

A corpus-based approach to cognitive semantic analysis

Stefan Th. Gries and Dagmar Divjak*

1. Introduction

In this paper we will look into questions that concern what may be considered two of the central meaning relations in semantics, i.e. polysemy or the association of multiple meanings with one form and synonymy, i.e. the association of one meaning with multiple forms.

In the domain of polysemy, cognitive semanticists typically face issues which center on the questions of how to determine whether two usage events are sufficiently similar to be considered instantiations of a single sense and how to establish the prototypicality of a sense/several senses; we adopt Evans's (2005: 33, n. 2) definition of sense as those meanings which have achieved conventionalization and are instantiated in semantic memory. In the domain of near synonymy, semanticists need to uncover among other things what syntactic, semantic and/or pragmatic differences there are between near synonyms and what the semantic and/or functional relation is between near synonyms in a semantic space. In order to solve these problems they need to be able to measure the degree of similarity between senses and/or words and to decide how and where to connect a sense/word to another sense/word in a network.

Several solutions to these problems have been put forward in the literature, in particular for polysemy-related issues. One such solution for polysemy-related issues is the full-specification approach inspired by Lakoff and his collaborators (cf. e.g. Norvig and Lakoff 1987; Lakoff 1987) where minimal perceived differences between usage events constitute different senses and image schemas. Related to this is Kreitzer's (1997) partial-specification approach where information from three different levels of schematization – the so-called component, relational, and integrative levels – is integrated, yet minimally different usage events need not constitute different senses. Both of these approaches suffer from methodological inadequacies and representational problems, however. As for the former approach, information provided by the context the word under study occurs in is not taken into account (cf. Sandra and Rice 1995; Tyler and Evans 2001), there is no

* At the time of writing, Dagmar Divjak was a Postdoctoral Research Fellow of the FWO-Vlaanderen (Belgium), working in the Research Unit QLVL at the KU Leuven (Belgium). The financial support of the FWO is gratefully acknowledged.

method for identifying how the primary sense has developed, and empirical support for fine-grained semantic distinctions is not provided. As for the latter approach, problems relate to the vagueness of the representation and the lack of clarity concerning the status of the proposed networks.

Contrary to the above-mentioned studies, Sandra and Rice (1995) and Rice (1996) measure the similarity of senses using a variety of experimental methods such as off-line sentence sorting followed by hierarchical cluster analyses, off-line sentence similarity judgments, on-line acceptability judgments and sentence generation. While this experimental approach is certainly more objective than introspective approaches, it is also a bit problematic. First, it remains unclear to what degree sentential context rather than the prepositions under investigation influence the subjects' sorting style (cf. Klein and Murphy 2001, 2002) as does the influence of methodological choices on the clustering. Second, the questions remain whether subjects use the same cognitive strategies for conscious off-line classification as for subconscious on-line production (a general problem of experimental approaches) and whether conscious off-line classification reflects the patterns underlying mental representation.

More recently the principled-polysemy approach was introduced by Tyler and Evans (2001). Tyler and Evans argue that previous research on polysemy lacks a constrained approach to distinguishing senses. For example, in their work on *over* they propose that a distinct sense of *over* should be posited if and only if the meaning of *over* in one utterance (i) involves a different spatial configuration from *over*'s use in another utterance and (ii) cannot be inferred from encyclopedic knowledge and/or context. In later work (on *time*) within the same framework, Evans (2005:41) introduces three criteria, which we quote here in detail because we will return to them later:

i. a **meaning** criterion: a distinct sense must contain additional meaning compared to other already established senses;
ii. a **concept elaboration** criterion: a distinct sense will feature unique or highly distinct patterns of concept elaboration [...] as in the lexical choices signaled by patterns of modification [...] or in the verb phrase which complements the noun phrase [...]. I assume that syntagmatic relations of this kind follow from semantic/conceptual considerations (see Croft's 2001 discussion of what he terms collocational dependencies);
iii. a **grammatical** criterion: a distinct sense "may manifest unique or highly distinct structural dependencies. That is, it may occur in unique grammatical constructions".

Although the last two criteria are in fact predictions about distributional patterns of the words under study, so far the proponents of the principled-polysemy approach have not utilized corpus data.

The second major question we raised above, namely how to determine the prototypical sense(s) of a word, has been an issue in polysemy ever since the first cognitive-linguistic analyses appeared. A variety of criteria has been proposed to isolate the prototypical sense (cf. e.g. Rice 1996: 145–146; Tyler and Evans 2001: Section 3.3; Evans 2005: Section 2.2.3) and the following is a non-exhaustive list of such criteria: asymmetrical judgments of goodness or similarity; ease of elicitation; gradation within the category; diachronically earliest sense; centrality/predominance in the semantic network; use in composite forms;

frequency of occurrence etc. Unfortunately, it remains unclear whether all criteria can be applied to all kinds of words and sometimes the proposed criteria make conflicting or counter-intuitive predictions (cf. Corston-Oliver 2001; Divjak and Gries 2006; Gries 2006). We admit, though, that this is a risk of all multifactorial approaches rather than a problem of any one particular study mentioned above.

Although near synonymy constitutes, in a sense, the opposite of polysemy, it has received relatively little attention in recent years. Within cognitive linguistics but a few studies have been devoted to the phenomenon (Geeraerts 1985; Mondry and Taylor 1992; Taylor 2003); this is likewise the case within western linguistics in general (Cruse 1986 being the exception). Surprisingly, the studies that have been carried out within the cognitive linguistic framework do utilize non-elicited material, yet the illustrative use of the corpus data makes them but mere forerunners of the corpus-based approach we will introduce below (see also Divjak 2004).

To sum up, in spite of the prominence the term 'usage-based' currently enjoys in cognitive-linguistic publications and in spite of the fact that some approaches explicitly couch their criteria in corpus-linguistic terms, there are few truly corpus-based approaches to polysemy and near synonymy. One laudable exception is the largely corpus-based approach of Kishner and Gibbs (1996) to *just* (as well as Gibbs and Matlock 2001 on *make*) which anticipated much of the above mentioned proposals by Evans (2005). Gibbs and colleagues investigate R1 collocates and colligations, correlating different senses with collocations and colligations.[1] Their "findings suggest the need to incorporate information about [...] lexico-grammatical constructions in drawing links between different senses of a polysemous word" (Gibbs and Matlock 2001: 234). Unfortunately, these studies do not fully utilize the potential of corpus data: citations in corpus data have more to offer than just individual collocations and colligations, and restricting the analysis to R1 collocates is a heuristic that is blind to syntactic structure (cf. points of critique also raised in collostructional analysis; cf. Stefanowitsch and Gries 2003; cf. Divjak 2006).

Work in corpus linguistics, on the other hand, has exploited the potential of corpus data more fully. Such studies start out from the self-evident statement that corpus data provide distributional frequencies. The assumption then is that distributional similarity reflects, or is indicative of, functional similarity, the understanding of functional similarity being rather broad, i.e. encompassing semantic, discourse-pragmatic, and other functions a particular expression can take on. Against this background, Atkins's (1987) study on *danger* involves collocate analysis from L7 to R7, colligations, part of speech (POS) characteristics of the head word, and all the collocations/colligations correlating (probabilistically or perfectly) with a particular sense are referred to as an ID tag. Also, Hanks's paper (1996) on *urge* involves collocate and colligation analysis. He argues that "the semantics

1. The term *collocation* encompasses both the probabilistic co-occurrence of word forms (e.g. *different to* vs *different than*) as well as the absolute frozenness of expressions (e.g. *by and large*). *Collocations* are thus co-occurrences of words which are referred to as *collocates*; often, the letters L (for left) and R (for right) are used together with a number to refer to the position of one collocate with respect to the head word (e.g. R1 meaning 'the first collocate to the right'). The term *colligations* refers to the co-occurrence of word forms with grammatical phenomena (e.g. the preference of *consequence* to occur as a complement and with an indefinite article).

of a verb are determined by the totality of its complementation patterns" (1996: 77), where a set of coarse complementation patterns and semantic roles of a word is referred to as a behavioral profile. Unfortunately, neither Atkins nor Hanks provides conclusive evidence concerning the predictive power of the ID tags investigated. In addition, much of the method of analysis remains to be fleshed out and lacks quantitative sophistication.

In other words, while interesting studies have been conducted, semantic analyses in the area of polysemy and near-synonymy have often been based on introspective data. This makes them not only empirically problematic, but it likewise prevents the development of a rigorous, quantifiable, and objectively comparable methodology. Corpus-based or computational-linguistic studies, on the other hand, do introduce methodological rigor, yet, they are rather limited as they treat words with few different senses or focus on small sets of semantically similar words (*almost* vs. *nearly*, *high* vs. *tall*, *between* vs. *through*). In addition, they use data that constitute impoverished subsets of what is actually available: basing a semantic analysis of words solely on collocates in one sequentially defined slot means both seriously limiting the data taken into consideration and disregarding the syntactic structure of the clause under investigation. Thirdly, the databases used in computational linguistic research may be noisy or skewed given that such studies often rely on (semi-) automatic preprocessing tools.

In this paper, we will argue in favor of a radically corpus-based approach to polysemy and near synonymy. The approach is *radically* corpus-based because we rely on the correlation between distributional patterns and functional characteristics to a much larger extent than most previous cognitive-linguistic work; we will clarify this statement below. We submit our approach is a worthwhile addition to the cognitive-semantic field: the notion *usage-based* is encountered more and more frequently – the principled-polysemy approach even makes explicit use of corpus-linguistic terms – and corpus-based approaches have a variety of advantages that include, but are not limited to, the following:

1. the criteria used are not based on traditional minimal pair acceptability tests, which often fail to account for more complex patterns (cf. Gries 2003: Section 2.6.2 for discussion of such shortcomings in the area of syntax);
2. judgments are not gathered in an introspective way that relies on implicit knowledge and thus makes it difficult to validate and replicate findings;
3. instead, corpora
   a. provide many instances rather than a few isolated judgments;
   b. provide data from natural settings rather than 'armchair' judgments or responses that potentially reflect experimentally-induced biases;
   c. provide co-occurrence data of many different kinds, i.e. not just those a particular researcher may consider important;
   d. and thus, allow for bottom-up identification of relevant distinctions as well as for a more comprehensive description than is typically provided.

In this study, we will introduce a methodology that aims to provide the best of both worlds, i.e. a precise, quantitative corpus-based approach that yields cognitive-linguistically relevant results.

2. Methods

Our method is based on two key concepts. One is the notion of *ID tag* as proposed by Atkins (1987). The other is Hanks's (1996) notion of *Behavioral Profile*, which we extend from being restricted to complementation patterns and roles to include a comprehensive inventory of elements co-occurring with a word within the confines of a simple clause or sentence in actual speech and writing.

Our approach hinges on the assumption that the words or senses investigated are part of a network of words/senses. In this network, elements which are similar to each other are connected in such a way that the strength of the connection reflects the likelihood that the elements display similar behavior in other linguistic subdomains. The corpus-based method we will introduce focuses on co-occurrence information of symbolic units since (i) the symbolic unit is considered the basic unit within a cognitive linguistic approach and (ii) co-occurrences of this type are most easily accessible for a corpus-based approach.

The method involves the following four steps:

i. the retrieval of (a representative random sample of) all instances of a word's lemma from a corpus;
ii. a (so far largely) manual analysis of many properties of the word forms (i.e. the annotation of the ID tags);
iii. the generation of a co-occurrence table;
iv. the evaluation of the table by means of exploratory and other statistical techniques.

The first three of these steps are concerned with data processing, and will be dealt with in Section 2.1. The fourth step is concerned with how the resulting data can be evaluated meaningfully; it will be covered in detail in Section 2.2.

2.1 Data processing

Let us go over the data processing steps in somewhat more detail. The first step involves using a concordancing program, a programming language (e.g. R or Perl), or a corpus interface to retrieve (a subset of) all hits for the lemmata of a word or words of interest.[2]

In the second step, all hits are annotated for the ID tags one wishes to include in the analysis (cf. Section 4 below for discussion) in such a way that the results of the annotation process can be imported into spreadsheet software at a later stage. The range of ID tags that can be used is vast since virtually every linguistic level of analysis can be included. Table 1 provides a summary of ID tags that have been used so far.[3]

2. Note in passing that we use lemmata in order to be able to investigate whether particular inflectional forms behave differently from others. However, nothing in particular hinges on this decision and one might just as well base the study on the frequencies resulting from combining all inflectional forms of a lemma (cf. Gries to appear for discussion).

3. This list of ID tags results from our work on English and Russian. It is not exhaustive as far as senses are concerned and could be extended with additional ID tags (from the same domains or others such as phonology or pragmatics) or with ID tags manifested in other languages.

Table 1. Selective overview of (kinds of) ID tags and their levels

Kind of ID tag	ID tag	Levels of ID tag
morphological	tense	present, past, future
	mode	infinitive, indicative, subjunctive, imperative, participle, gerund
	aspect	imperfective vs. perfective
	voice	active vs. passive
	number	singular vs. plural
	transitivity	intransitive, monotransitive, copular, complex transitive
syntactic	sentence type	declarative, exclamative, imperative, interrogative
	clause type	main vs. dependent
	type of dependent clause	adverbial, appositive, relative, zero-relative, zero-subordinator, etc.
semantic	semantic types of subjects, objects, etc.	concrete vs. abstract, animate (human, animal) vs. inanimate (event, phenomenon of nature, body part, organization/institution, speech/text) etc.
	countability of nouns	count vs. mass
	properties of the process denoted by the verb	physical actions, physical perception, communication, intellectual activities, emotions, etc.
	controllability of actions	high vs. medium vs. no controllability
	adverbial/PP modification (if present)	temporal, locative, etc.
	negation	present vs. absent, attached to which element
lexical	collocates in precisely-defined syntactic slots (i.e. collexemes)	$collocate_1$, $collocate_2$, …, $collocate_n$

Table 2. An excerpt from a co-occurrence table for *to run*

Citation	transitivity	morph. form	clause type	sense
Bert's now the priest who **runs** it	monotrans	present tense	depend	to manage
I will **run** out of money	intrans	infinitive	main	to lack
Troopers said the child **ran** into the path of a passing car	intrans	past tense	depend	to go very rapidly

The result of the second step is a table with co-occurrence information. In other words, each row contains one citation of the word in question, each column contains an ID tag and each cell contains the level of the ID tag for this citation. Table 2 contains an excerpt from the table used for the analysis of polysemous *run* in English (with examples from the ICE-GB). An analogous table for the investigation of near synonymous words would feature the near synonym in the last column (instead of the sense of a polysemous word).

In a third step, this table is prepared for quantitative analysis; this step consists of two phases. First, Table 2 is turned into a frequency table in a way that every row contains a level of an ID tag while every column contains a sense of the polysemous word

Table 3. Absolute co-occurrence frequencies of (levels of ID tags) and word senses

ID tag	level of ID tag	manage	lack	go very rapidly	
transitivity	intransitive	0	12	191	
	monotransitive	101	1	12	203
	copular	0	1	0	
	complex transitive	0	0	0	
morphological form	infinitive	25	1	43	
	present tense	15	5	11	
	present participle	23	4	54	203
	past tense	10	2	78	
	past participle	28	2	11	
	imperative	0	0	6	

Table 4. Relative co-occurrence frequencies of (levels of ID tags) and word senses

ID tag	level of ID tag	manage	lack		go very rapidly
transitivity	intransitive	0	0.8571		0.9409
	monotransitive	1	0.0714	sum to 1	0.0591
	copular	0	0.0714		0
	complex transitive	0	0		0
morphological form	infinitive	0.2475	0.0714		0.2118
	present tense	0.1485	0.3571		0.0542
	present participle	0.2277	0.2857	sum to 1	0.2660
	past tense	0.0990	0.1429		0.3842
	past participle	0.2772	0.1429		0.0542
	imperative	0.0000	0		0.0296

or one word of the set of near synonyms; consequently each cell in the table provides the frequency of occurrence of the ID tags with the word/sense (cf. Table 3). The summed frequencies within each ID tag must be the same: for the sense *go very rapidly* this means that the sum of transitivity related ID tags (191+12) equals the sum of ID tags that capture morphological form (43+11+54+78+11+6).

In order to compare senses that occur at different frequencies, the absolute frequencies from Table 3 need to be turned into relative frequencies (i.e. within ID tag percentages; cf. Table 4).

In a quantitative, narrow sense of the term, Table 3 and Table 4 form the *behavioral profile* for a word/sense. In other words, each sense of a word or each near synonym within a semantic domain is characterized by one co-occurrence vector of within-ID tag relative frequencies.[4] It is worth pointing out that this approach is compatible with at least two of the criteria of the principled-polysemy framework, namely the concept-elaboration criterion, positing distinct syntagmatic co-occurrence relations, and the grammatical criterion, positing distinct grammatical constructions. In fact, one could even say that our

4. Thus, the notion of behavioral profile is not related to the concept of profiling in cognitive grammar.

behavioral profile approach is based on taking these criteria and their manifestations as seriously as present-day corpora and efficiency demands allow. The following section will explain how behavioral profiles can be evaluated.

2.2 Evaluation

The vector-based behavioral profile can be subjected to a variety of quantitative approaches for further evaluation. There exist monofactorial and/or pairwise approaches as well as more comprehensive techniques that account for more complex multifactorial patterns. In Section 2.2.1, we will introduce some monofactorial methods, which will then be exemplified in more detail on the basis of the English verb *run* in Section 3.1. In Section 2.2.2, we will introduce a multifactorial cluster-analytic method, the application of that method to Russian verbs that express 'try' will be exemplified in Section 3.2.

2.2.1 *Monofactorial evaluation*

The most straightforward ways of analyzing behavioral profiles are looking at both token frequencies and type frequencies. Let us start with token frequencies. A useful first strategy is identifying in the corpus the most frequent senses of the word(s) one is investigating or the most frequent words within the semantic field studied. So far, our discussion has been non-committal with respect to the type of corpus investigated, but depending on the corpus the identification of the most frequently occurring word(s) or sense(s) may license different conclusions. In a general synchronic corpus, overall token frequency may be correlated with the degree of entrenchment of a word sense or of a word in a semantic field as well as its prototypicality (cf. Geeraerts 1988: 222; Winters 1990). In an acquisition corpus, tracking high percentages of senses and words across time and monitoring how they change over time may license conclusions about the ease of acquisition of senses and words as well as straightforward ways of semantic extension. In a diachronic corpus, the same procedure allows us to concentrate on the historical primacy of senses or words as well as on possible paths of extension and grammaticalization. While corpus-based work has been carried out in all of these areas, it typically takes a slightly more restricted stance in that the behavioral profiles entering into the analyses tend to be confined to many fewer ID tags than we propose.

While the inspection of frequencies is ultimately based on high *token* frequencies of particular ID tags, inspecting the *type* frequencies of ID tags is also revealing. Type frequencies should be 'normalized', i.e. the number of ID tags should be corrected against the overall frequency of occurrence of the sense or word (for instance, by dividing the number of observed ID tag types by the frequency of occurrence of that sense or word). The word senses or words with the highest number of non-zero values, i.e. the highest number of different ID tags, found in the behavioral profile correspond to unmarked senses or words since these senses/words exhibit the fewest restrictions concerning the range of ID tags applicable to them. Again, this may be an interesting finding in itself, as there is a positive though by no means absolute correlation between markedness and prototypicality (cf. Lakoff 1987: 60–61) which may be worth exploring. Yet, data of this type also allow the identification of exactly those cases where the co-occurrences of

senses/words and particular ID tags seem impossible, which in turn invites interesting semantic conclusions. Croft (1998: 169), for example, argues that disjoint syntactic-semantic distributions of otherwise similar senses support splitting senses as opposed to lumping them together.[5]

In addition, the distributional form in which the data come allows for more technical approaches from computational linguistics, where vectorized data underlie work on the semantic similarity of words, document clustering, and information retrieval (cf. Manning and Schütze 2000: Section 8.5). Moreover, the behavioral profile facilitates quantifying (and, thus, rank-ordering) senses or words in terms of their pairwise similarity (for more complex approaches, cf. Section 2.2.2 below); this goal can be achieved by computing any of several available similarity measures for vectors such as standard correlation coefficients, cosines, or other more complex indices. For example, network-inspired analyses of polysemous words require decisions as to where to locate senses in the network, and one way of approaching this issue is to first determine the highest pairwise similarities of the senses/words in question and then connect them to the senses/words they are most similar to.

One common characteristic of all of the above listed techniques is that they are monofactorial. That is, they are built either on vectors, i.e. one-dimensional distributions of percentages, or on pairwise similarities between vectors. However, the behavioral profile approach we are promoting here has more to offer and in the next section we will outline how multifactorial techniques can be brought to bear on the issues raised so far.

2.2.2 *Multifactorial evaluation*

There is quite a number of multifactorial techniques that could be applied to extract relevant information from behavioral profiles; we will restrict our attention to the exploratory technique of hierarchical agglomerative cluster analysis since it has been applied most frequently in related domains (cf. Manning and Schütze 2000: Chapter 14 for examples and discussion).[6] The kind of cluster analysis that we advocate can be seen as consisting of three different steps, which we will discuss in turn.

The first step of the analysis consists of the hierarchical agglomerative cluster analysis proper of the joint behavioral profiles under investigation. Hierarchical agglomerative cluster (HAC) analysis is a family of methods that aims at identifying and representing (dis)similarity relations between different items; cf. Kaufman and Rousseeuw (1990)

5. It may likewise be possible to use the distributional data for exploring the acquisition of senses/words in a way complementing the approach mentioned above: equally frequent senses/words may differ in terms of their co-occurrence restrictions. A viable question would then be whether the more widely distributed senses/words give rise to extension of the category earlier than the more restricted ones. A similar logic applies to the case of diachronic corpora; cf. Bybee and Thompson (1997) for a pertinent discussion on type vs. token frequencies.

6. Techniques other than cluster analyses that can be applied to the kind of data discussed are singular value decomposition techniques (such as factor analyses or LSA), techniques for the multidimensional analysis of frequency tables (such as loglinear analysis or configural frequency analysis) and tree-based classification methods.

for a general discussion of clustering. Usually, clustering is performed on the basis of variables that characterize the items or on the basis of a (dis)similarity matrix of the items. In the TRY case, 1,585 corpus extractions that include examples for all nine near-synonymous verbs were tagged for 87 variables, i.e. our ID tags (a selection of which is shown in Table 1). Assigning ID tags to extractions resulted in the dataset represented in Table 4 above. Table 4 needs to be turned into a similarity/dissimilarity matrix, however, which can be done by means of a suitable similarity/dissimilarity measure. Since there are several measures available which differ along one or more parameters and thus may yield different cluster solutions, it is impossible to recommend any one specific measure: what is most suitable in one case (or with one set of assumptions one has about the data) may not work in another. It is probably fair to say that Euclidean distances (or squared Euclidean distances if one wants to 'punish' outliers) are among the most widely used measures in linguistic analyses.

Once the similarity/dissimilarity matrix has been generated, an amalgamation strategy has to be selected. An amalgamation strategy is an algorithm that defines how the elements that need to be clustered will be joined together on the basis of the variables or ID tags that they were inspected for. Again, the same caveats apply as for the generation of the similarity/dissimilarity matrix. One of the most widely used amalgamation strategies is Ward's rule: it is conceptually similar to the logic underlying analysis of variance and typically yields moderately sized clusters.[7]

The result of such an analysis is a hierarchical tree diagram representing, in the ideal case, several relatively easily distinguishable clusters that are characterized by high within-cluster similarity and low between-cluster similarity. Often, the information gleaned from such a diagram is revealing in itself since the diagram summarizes conveniently what a human analyst could hardly discern given the complexity of a multifactorial data set.

The second step of the analysis consists of a detailed analysis of the clustering solution which (i) assesses the 'cleanliness' of the tree diagram and (ii) focuses on precisely those kinds of similarity that emerge most clearly from the tree diagram: between-cluster similarity and within-cluster similarity (cf. Backhaus et al. 2003: Chapter 8). As to the former, by a variant of the *F*-test also used in analyses of variance, it is possible to determine how homogenous the obtained clusters are. Obviously, the more homogenous the clusters are, the easier the interpretation of the between-cluster differences will be. As to the latter, it is possible to use *t*-values to determine which of the ID tags used reflect between-cluster differences best. More specifically, one can compute a *t*-value for each ID tag for each cluster such that a positive/negative *t*-value of an ID tag for a cluster indicates that this ID tag is respectively over-represented or under-represented in that cluster. This way, it is, for instance, possible to identify ID tags that have a positive *t*-value in one cluster and negative values in all other clusters, thus revealing the scales of variation that matter most for the clustering solution.

7. An alternative possibility is the choice of a phylogenetic clustering algorithm (cf. Felsenstein 2005 for an implementation), which does not require all elements that need to be clustered to be merged into a single root.

The third and final step consists of a similarly detailed analysis of the within-cluster differences. The fact that a cluster analysis has grouped together particular senses/words does not necessarily imply that these senses or words are identical or even highly similar – it only shows that these senses/words are more similar to each other than they are to the rest of the senses/words investigated. By means of standardized *z*-scores, one can tease apart the difference between otherwise highly similar senses/words and shed light on what the internal structure of a cluster looks like.

While the discussion has been relatively abstract so far, we will now present several examples to illustrate how the methods introduced above can be put to use.

3. Examples

In this section, we will discuss examples from a case study on an extremely polysemous English verb (Section 3.1) and from a case study of nine near synonymous Russian verbs (Section 3.2).

3.1 Polysemy: The English verb *run*

The examples to be discussed in this section are taken from Gries (2006) that deals with the highly polysemous English verb *run*.[8] The analysis is carried out using 815 citations of the verb lemma *run* from two corpora; each citation was coded for the senses they instantiate within their respective contexts as well as for 252 ID tags of the types given in Table 1; many of the ID tags in this study code the presence/absence of particular collocates.

Let us begin with the issue of how one-dimensional vectors (frequency distributions) can be exploited to address the question of prototypical word senses, an issue where corpus data can be applied in a versatile way. In this case, the corpus data clearly single out one sense, namely the sense 'fast pedestrian motion'. This is the sense that is

- diachronically primary: together with *'flow'* it is the earliest attested sense;
- diachronically primary for the zero-derived noun *run*;
- synchronically most frequent in the analyzed corpora;
- synchronically most frequent for the zero-derived noun in the analyzed corpora;
- acquisitionally primary in the sense of being acquired earliest;
- acquisitionally most frequent (counts from data for Abe, Adam, Eve, Naomi, Nina, Peter, and Sarah from the CHILDES database; cf. MacWhinney 2000);
- combinatorially least constrained in the analyzed corpora (given its number of ID tags normalized against frequency of occurrence).

8. Cf. Langacker (1988) and Taylor (1996, 2000) for cognitive-linguistic but methodologically very different studies of the verb *run*.

Vectors can likewise be used to identify disjoint distributions, as the examples of 'fast pedestrian motion' and 'escape' show. Applying Croft's (1998) logic, for example, one would not consider the senses instantiated in (1a) and (1b) as different merely because their PPs highlight different landmarks. This is so because there are also examples like (2) in which the two kinds of PPs – SOURCE and GOAL – co-occur, showing that the distribution of the PPs is not disjoint.

(1) a. and we ran back [$_{\text{GOAL}}$ to my car]
b. Durkin and Calhoun came running [$_{\text{SOURCE}}$ from the post]

(2) I ran [$_{\text{SOURCE}}$ from the Archive studio] [$_{\text{GOAL}}$ to the Start The Week studio]

However, there are other senses, intuitively very similar, which are likely candidates for being lumped together. For example, there are two senses that could both be paraphrased as 'escape', but one of them involves moving away from something undesirable while the other involves moving away to engage in a romantic relationship. Interestingly, the former (see (3)) is attested with a SOURCE but not with a comitative argument whereas the latter (see (4)) is attested with a comitative but not with a SOURCE although both unattested combinations are conceivable.

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

(4) If Adelia had felt about someone as H. felt about C., would she have run away [$_{\text{COMITATIVE}}$ with him]?

While the results of a corpus-based application of the criterion of disjoint distribution are certainly dependent on sample sizes, they indicate – in the absence of evidence to the contrary – that the two 'escape' senses should not be lumped together. Once it has been decided to keep these senses separate the question arises of where to connect them to the rest of the network. One possible point of connection would be the sense of 'fast pedestrian motion'. Yet, not all the instances of the 'escape' senses imply fast pedestrian motion: some merely imply 'fast motion' or only 'motion'. 'Motion' would therefore also be a plausible candidate sense for the connection. This issue can be solved by making use of the information contained in the behavioral profile for each sense. Pearson product moment correlations were computed for all pairs of senses in order to determine the average correlation of all senses but also to find out which of the three candidate senses are most similar to the two 'escape' senses that need to be connected. While the overall average correlation (after Fisher Z transformation) was moderate ($r = 0.545$), the average correlation of the two 'escape' senses and the three 'motion' senses was considerably higher ($r = 0.848$), supporting the intuition that these senses are in fact closely related, at least much more than they are related to the multitude of other senses that *run* can have. When the question of where to attach the two 'escape' senses was investigated using a smaller set of ID tags (omitting collocation-based ID tags lest individual collocates distort the picture), a surprisingly clear answer emerged. The two 'escape' senses were significantly more similar to 'fast pedestrian motion' than to the other two senses, which in turn did not differ significantly from each other. This result provides evidence for attaching the two 'escape' senses to the prototypical sense as opposed to the two slightly more general senses.[9]

So far the examples presented involved only monofactorial data (for considerations of space, the cluster-analytic results presented in Gries 2006 are not discussed here). The following section will provide detailed exemplification of how cluster analyses and their follow-up investigation can be useful for the lexical semanticist.

3.2 Near synonymy: Russian verbs meaning *try*

In this section, based on Divjak and Gries (2006), we show how clustering behavioral profiles and evaluating clusters and verbs in terms of *t*-values and *z*-scores provide us with scales of variation for describing and distinguishing near synonyms in a fine-grained lexical semantic analysis. Divjak and Gries (2006) analyze 1,585 sentences each containing one out of nine Russian verbs that, in combination with an infinitive, express *try*. Since the verbs in question differ strongly in terms of their frequencies, the sentences were culled from several sources, keeping the genre constant: the Amsterdam corpus, the Russian National Corpus, and the WWW (cf. Divjak and Gries 2006: 54, note 6 for detailed discussion of the sampling procedure); Table 5 sketches the composition of the data set.

All 1,585 sentences were annotated for 87 ID tags; as a result, for each of the nine verbs a behavioral profile vector was obtained of the sort exemplified in Table 4. This dataset was analyzed using a hierarchical agglomerative cluster analysis (similarity metric: Canberra; amalgamation strategy: Ward), resulting in the dendrogram presented in Figure 1. The tree plot shows what is similar and what is different: items that are clustered or amalgamated early are similar, and items that are amalgamated late are rather dissimilar.

For example, it is obvious that *pytat'sja* and *starat'sja* are much more similar to each other than, say, *probovat'* and *norovit'*, which are only linked in the last overarching cluster. At the same time, the plot gives an indication of how independent the clusters are: the larger the distance between different points of amalgamation, the more autonomous the earlier verb/cluster is from the verb/cluster with which it is merged later. In the present case, the plot clearly consists of three clusters.

Table 5. Composition of the dataset analyzed in Divjak and Gries (2006)

Verb	N (AC/RNC/Web)	Verb	N (AC/RNC/Web)
probovat'	246 / – / –	*poryvat'sja*	31 / 88 / –
pytat'sja	247 / – / –	*tščit'sja*	21 / 30 / 21
starat'sja	248 / – / –	*pyžit'sja*	– / – / 98
silit'sja	57 / 185 / –	*tužit'sja*	– / – / 53
norovit'	112 / 148 / –		

9. Of course, this method is not restricted to cases where one sense needs to be attached to only one other sense. In cases where multiple attachments are desired, the correlations can still be used to rank or delimit the candidate set of senses to which another sense can be reasonably attached. Also, nothing hinges on the choice of the Pearson product moment correlation: as indicated above, other measures could be employed; in this particular case, the cosine measure was also tested and yielded the same conclusions.

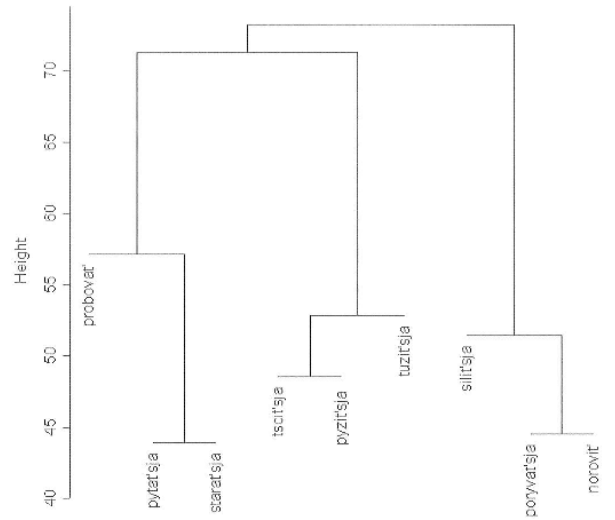

Figure 1. Dendrogram for tentative verbs in Russian

A cognitive approach to language and particularly to the incorporation of knowledge about human categorization mechanisms into linguistics provides interesting perspectives for a unified interpretation of the data. On the cognitive linguistic approach, (linguistic) categories may exhibit prototype effects and instantiate radial networks of related expressions with semantically motivated connections (Lakoff 1987: Chapter 6).[10] In order to investigate the nature of the three categories suggested by the dendrogram more thoroughly, between- and within-cluster similarities and differences were inspected using *t*-values and *z*-scores (cf. above); limitations of space permit only a selection of the results to be discussed.

The first cluster groups together [[*pytat'sja* and *starat'sja*] and *probovat'*]. All verbs in this cluster are more easily used in the main clause ($t = 0.821$) than verbs from the other two clusters. Although all three verbs exist in the imperfective and perfective aspect and do occur in both aspects, variables that include reference to the perfective aspect (i.e. refer to past and future events) are three times more frequent in the top 25 *t*-scores that are positive for this cluster and negative for other clusters (*t*-values range from 0.667 to 1.201). In addition, the infinitive that follows the tentative verb is more often negated ($t = 0.702$) and expresses physical activities ($t = 0.599$), events that are figurative extensions of motion events ($t = 0.465$) or involve setting a theme/patient into motion ($t = 0.4$). Finally, strongly

10. Although the HAC dendrogram presented in Figure 1 can be manually transformed into a radial network representation, Divjak and Gries (2006) backed up their results by analyzing the distance matrix resulting from the behavioral profiles using a phylogenetic clustering algorithm, the Fitch program from the PHYLIP package (Felsenstein 2005). The results were for all practical purposes identical; cf. Divjak and Gries (2006: Section 3) for discussion.

attracted optional collocates express that the subject got permission to carry out the infinitive action (using *pust'*, $t = 1.008$), that the attempt was untimely brought to a halt (with *bylo*, $t = 0.982$), that the subject was exhorted to undertake an attempt ($t = 0.832$) and that the intensity with which the attempt was carried out was reduced ($t = 0.667$).

In the middle, there is a cluster that unites the imperfective verbs [*tščit'sja* and *pyžit'sja* and *tužit'sja*]. All three verbs lack a perfective counterpart and prefer the present tense more than verbs in the two other clusters ($t = 1.047$ for present tense with a perfective infinitive and $t = 0.711$ for the present tense followed by an imperfective infinitive). Among the most strongly represented variables we encounter the verbs' compatibility with inanimate subjects, both concrete and abstract (*t* ranges from 1.108 to 1.276), as well as with groups or institutions ($t = 1.297$). Actions expressed by the infinitive are physical ($t = 0.176$), affect a theme/patient ($t = 0.352$), are metaphorical extensions of physical actions ($t = 0.999$), or physical actions affecting a theme/patient ($t = 0.175$). Focus is on the vainness ($t = 0.962$ for vainness combined with intensity) of the durative effort ($t = 0.750$ for duration adverbs).

The third cluster, amalgamated last into the overarching cluster, consists of [[*norovit'* and *poryvat'sja*] and *silit'sja*]. These verbs prefer to occur as participles (*t*'s range from 0.632 to 1.214). The infinitive actions that are attempted express a type of physical motion ($t = 0.924$) that is often not controllable ($t = 0.548$). The action can be carried out by an inanimate subject ($t = 0.809$ for phenomena of nature and $t = 0.774$ for bodyparts) and are often repeated (*t* ranges from 0.678 to 1.092). If the attempt remains unsuccessful, both external ($t = 0.627$) and internal ($t = 0.429$) reasons are given for the failure.

Apart from between-cluster differences that are revealed by means of *t*-scores, *z*-values make within-cluster similarities and differences visible. As an illustration, let us look at the three most frequently used verbs, i.e. the verbs in the first cluster [[*pytat'sja* and *starat'sja*] and *probovat'*]. The two verbs that are clustered first, *pytat'sja* and *starat'sja*, resemble each other to a large extent, yet a close inspection of their distributional properties reveals that *pytat'sja* is more strongly attracted to occurring in the past tense (with *z*'s ranging from 1.092 to 1.155, all with perfective infinitives) whereas *starat'sja* is relatively more often found in the present tense ($z = 1.153$ with imperfective infinitives). *Pytat'sja* is not particularly attracted to weakly controllable actions ($z = -1.097$) whereas *starat'sja* avoids controllable actions ($z = -1.049$). *Starat'sja* combines, among other things, with passive perception verbs ($z = 1.134$), whereas *pytat'sja* goes well with mental activities ($z = 1.139$). *Starat'sja* is frequently found with a negated infinitive ($z = 1.151$), thus indicating that the subject is avoiding an event that might take place. Easiest to interpret is the verbs' preference for different adverbs: *starat'sja* is most strongly characterized by adverbs that express repetitive duration (*vsë vremja*, $z = 1.155$), reduced intensity ($z = 1.155$), and intensity ($z = 1.101$), whereas *pytat'sja* prefers repetition ($z = 1.111$). In other words, if one has already applied *pytat'sja* without success, a possible way to achieve the desired result despite the initial failure is by using what is encoded in *starat'sja* (cf. (5)).

(5) Он убрал Мазера и Леоновича, **постарается то же проделать с Казаковым** *(уже пытался)*, и весьма возможно, с Соя-Серко.
[Ф. Незнанский, Ярмарка в Сокольниках]
'He took away Mazer and Leonovič, is trying (hard) [starat'sja] to do the same with Kazakov (he has already tried [pytat'sja]), and it is very likely, with Soja-Serko.'

Added to [*pytat'sja* and *starat'sja*] is the verb *probovat'* that is rather dissimilar. This verb occurs preferably in a main clause ($z = 1.127$), and is not typically found in declarative clauses ($z = -1.148$). Tags that refer to perfective aspect receive the highest z-scores for [*probovat'*], ranging from 1.003 to 1.155. Although all three verbs in this cluster have a perfective counterpart formed by means of the delimitative prefix *po-*, *po/probovat'* significantly prefers the perfective aspect in 74.8% of all examples while *pytat'sja* and *starat'sja*, by contrast, significantly prefer the imperfective aspect, i.e. in 79.6% and 83% of all cases respectively ($\chi^2 = 222.72$; $df = 2$; $p < 0.001$, Cramer's $V = 0.548$). Related to the more frequent use of perfective forms is the possibility of locating the attempt in the future ($z = 1.003$ for combinations with imperfective infinitives and $z = 1.044$ with perfective infinitives), as well as a considerable relative dispreference for the present tense (z's ranges from = –0.632 to –1.154). Finally, *probovat'* is the only verb that is often found in the imperative mode (with z's ranging from 1.092 to 1.134). In interpretive terms, the node [*probovat'*] uses the perfective to present each try as a completed entity. This allows the subject to change method or strategy between attempts, which might be what makes this verb resemble experiments (cf. Wierzbicka 1988: 309; Apresjan et al. 1999: 304). An experimental attempt is also demanded more easily from another person than attempts that require long and/or intense effort, hence the higher frequency of the imperative and attraction of exhortative particles ($z = 1.121$). Failure can be attributed to internal and external factors alike (4.9%, $z = 1.155$ and 11%, $z = 1.151$). In all, *probovat'* seems to be less intensive than *pytat'sja* (and *starat'sja*), as example (6) shows.

(6) Бим *уже* **пробовал на нее наступить**, но пока еще так, *немножко – только пробовал*. [Г. Трупольский. Белый Бим черное ухо]
'Bim had already tried [probovat'] to step on her, but just like that, a little bit, he had only tried [probovat'].'

The multifactorial evaluation we propose comprises a set of both exploratory and hypothesis-testing statistical techniques for analyzing corpus-based behavioral profiles. We have illustrated how, on the basis of these results, the internal structure of a cluster of near synonymous verbs can be laid bare and the verbs in those clusters can be compared.

4. Conclusion

We hope to have shown that behavioral profiles and the proposed methods for their evaluation are valuable for the analysis of polysemous and near synonymous items in particular as well as for lexical-semantic research in general. Moreover, behavioral profiles provide an ideal starting point for research concerning interfaces between different levels of linguistic analysis, e.g. the syntax-lexis interface, and offer a wealth of usage-based evidence for cognitive linguistic theorizing concerning network representations, prototypicality of senses, sense-distinctions and the polysemy-homonymy discussion to name but a few. In addition, results of this type may also be relevant for researchers from neighboring disciplines, such as psycholinguistics: behavioral profiles can be used in formulating and evaluating hypotheses concerning the interaction between grammar and lexicon in language acquisition as well as with respect to the mental reality of radial categories (cf. Divjak and Gries 2008). Conveniently, a program for converting annotated data into behavioral-profile vectors and computing cluster-analytic statistics is now available (cf. Gries 2008).

Our plea for a corpus-based approach does not imply adherence to a fully automated approach, however. At present there is no reliable way for assigning (many) ID tags automatically and neither can a machine interpret statistical results. Although human intervention rules out complete objectivity, we do claim that our methodology is more objective than many others currently available. The proposed approach requires all information entering into the analysis to be made explicit: it is necessary to define and operationalize every ID tag since it is only through frequency counts of ID tags that information can be included. In other words, our method helps to minimize the share of subjective, implicit knowledge. In addition, while the choice of ID tags to be included in the analysis and the subsequent interpretation of the results contain elements of subjectivity – as does, if to a lesser degree, the annotation/coding of the dataset – a substantial part of the analysis is entirely objective. For example, an analyst cannot simply select parameters or ID tags for interpretation *ad libitum*, but is strongly constrained by the statistical results which were arrived at in an objective and replicable way (a hierarchical agglomerative cluster analysis can be defined precisely in terms of its mathematical settings). Thus, if, say, a t-score does not differentiate (significantly) between clusters, the analyst cannot belabor its importance however much his theoretical commitment would require him to. For these reasons we submit that the behavioral profile approach as outlined above is an improvement over many other methodological tools in the domain of lexical semantics in general and cognitive lexical semantics in particular.

References

Apresjan, J. D., Boguslavskaja, O. J., Levontina, I. V., Uryson, E. V., Glovinskaja, M. J. & Krylova, T. V. 1999. *Новый объяснительный словарь синонимов русского языка. Volume I.* Moskva: Škola "Jazyki Russkoj Kul'tury".

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.

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. 2003. *Multivariate Analysemethoden: Eine Anwendungsorientierte Einführung*. Berlin: Springer.

Bybee, J. & Thompson, S. 1997. Three frequency effects in syntax. *Proceedings of the Twenty-third Annual Meeting of the Berkeley Linguistics Society*: 378–388.

Corston-Oliver, M. 2001. Central meanings of polysemous prepositions: Challenging the assumptions. Poster presented at the ICLC 2001, University of California at Santa Barbara.

Croft, W. 1998. Linguistic evidence and mental representations. *Cognitive Linguistics* 9 (2): 151–173.

Croft, W. 2001. *Radical Construction Grammar*. Oxford: Oxford University Press.

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

Divjak, D. 2004. Degrees of Verb Integration. Conceptualizing and Categorizing Events in Russian. PhD Dissertation, K.U. Leuven (Belgium).
Divjak, D. 2006. Ways of intending: Delineating and structuring near synonyms. In *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*, St. Th. Gries & A. Stefanowitsch (eds), 19–56. Berlin: Mouton de Gruyter.
Divjak, D. & Gries, St. Th. 2006. Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2 (1): 23–60.
Divjak, D. & Gries, St. Th. 2008. Clusters in the mind? Converging evidence from near synonymy in Russian. *The Mental Lexicon* 3 (2): 188–212.
Evans, V. 2005. The meaning of *time*: Polysemy, the lexicon and conceptual structure. *Journal of Linguistics* 41 (1): 33–75.
Felsenstein, J. 2005. *Phylogeny Inference Package PHYLIP 3.65*, http://www.phylip.com.
Geeraerts, D. 1985. Preponderantieverschillen bij bijna-synoniemen. *De Nieuwe Taalgids* 78: 18–27.
Geeraerts, D. 1988. Where does prototypicality come from? In *Topics in Cognitive Linguistics*, B. Rudzka-Ostyn (ed.), 207–229. Amsterdam: John Benjamins.
Gibbs, R. W. & Matlock, T. 2001. Psycholinguistic perspectives on polysemy. In *Polysemy in Cognitive Linguistics*, H. Cuyckens & B. Zawada (eds), 213–239. Amsterdam: John Benjamins.
Gries, St. Th. 2003. *Multifactorial Analysis in Corpus Linguistics: The Case of Particle Placement*. London: Continuum.
Gries, St. Th. 2006. Corpus-based methods and cognitive semantics: The many meanings of *to run*. In *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*, St. Th. Gries & A. Stefanowitsch (eds), 57–99. Berlin: Mouton de Gruyter.
Gries, Stefan Th. 2008. BP. A program for R (for Windows).
Gries, St. Th. To Appear. Corpus data in usage-based linguistics: What's the right degree of granularity for the analysis of argument structure constructions? In *Expanding Cognitive Linguistic Horizons*, M. Brdar, M. Ž. Fuchs & St. Th. Gries (eds).
Hanks, P. 1996. Contextual dependency and lexical sets. *International Journal of Corpus Linguistics* 1 (1): 75–98.
Kaufman, L. & Rousseeuw, P. J. 1990. *Finding Groups in Data*. New York: John Wiley.
Kishner, J. M. & Gibbs, R. W. 1996. How *just* gets its meanings: Polysemy and context in psychological semantics. *Language and Speech* 39 (1): 19–36.
Klein, D. E. & Murphy, G. L. 2001. The representation of polysemous words. *Journal of Memory and Language* 45 (2): 259–282.
Klein, D. E. & Murphy, G. L. 2002. Paper has been my ruin: Conceptual relations of polysemous senses. *Journal of Memory and Language* 47 (4): 548–570.
Kreitzer, A. 1997. Multiple levels of schematization: a study in the conceptualization of space. *Cognitive Linguistics* 8 (4): 291–325.
Lakoff, G. 1987. *Women, Fire, and Dangerous Things*. Chicago: University of Chicago Press.
Langacker, R. W. 1988: A usage-based model. In *Topics in Cognitive Linguistics*, B. Rudzka-Ostyn (ed.), 127–161. Amsterdam: John Benjamins.
MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum.
Manning, C. D. & Schütze, H. 2000. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: M.I.T. Press.
Mondry, H. & Taylor, J. R. 1992. On lying in Russian. *Language and Communication* 12 (2): 133–143.
Norvig, P. & Lakoff, G. 1987. Taking: a study in lexical network theory. In *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*: 195–206.
Rice, S. 1996. Prepositional prototypes. In *The Construal of Space in Language and Thought*, M. Pütz & R. Dirven (eds), 135–165. Berlin: Mouton de Gruyter.
Sandra, D. & Rice, S. 1995. Network analyses of prepositional meaning: Mirroring whose mind – the linguist's or the language user's? *Cognitive Linguistics* 6 (1): 89–130.

Stefanowitsch, A. & Gries, St. Th. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8 (2): 209–243.
Taylor, J. R. 1996. On running and jogging. *Cognitive Linguistics* 7 (1): 21–34.
Taylor, J. R. 2000. Approaches to word meaning: the network model (Langacker) and the two-level model (Bierwisch). In *The Lexicon-Encyclopedia Interface*, B. Peeters (ed.), 113–141. Amsterdam: Elsevier.
Taylor, J. R. 2003. Near synonyms as co-extensive categories: *high* and *tall* revisited. *Language Sciences* 25 (3): 263–284.
Tyler, A. & Evans, V. 2001. Reconsidering prepositional polysemy networks: The case of *over*. *Language* 77 (4): 724–765.
Wierzbicka, A. 1988. *The Semantics of Grammar*. Amsterdam: John Benjamins.
Winters, M. 1990. Toward a theory of syntactic prototypes. In *Meanings and Prototypes: Studies in Linguistic Categorization*, S. L. Tsohatzidis. (ed.), 285–306. London: Routledge.