Dagmar Divjak and Stefan Th. Gries

University of Sheffield and University of California at Santa Barbara

CORPUS-BASED COGNITIVE SEMANTICS A CONTRASTIVE STUDY OF PHASAL VERBS IN ENGLISH AND RUSSIAN[1]

Abstract: In this paper we will present a corpus-based cognitive-semantic analysis of five verbs that express 'begin' in English and Russian, i.e. *begin*, *start*, *načinat'/načat'*. *načinat'sja/načat'sja* and *stat'*. On the basis of a quantitative analysis of data extracted from the ICE-GB and the Uppsala Corpus we conclude that the prototype for each verb and each set of verbs in each language revolves around different characteristics altogether: the difference between *begin* and *start* is lexical in nature, that between *načinat'/načat'* and *stat'* can be described as aspectual, whereas the differences between *načinat'/načat'* and *načinat'sja/načat'sja* involve argument structure. Because these dissimilarities are of an entirely different order, they can only be picked up if a methodology is used that adequately captures the multivariate nature of the phenomenon. The Behavioral Profiling approach we have developed and apply here does exactly that.

Keywords: Behavioral Profiling, ID tag, verb sense, complementation, near-synonymy, polysemy, prototype, English, Russian.

1. Introduction

Ever since the emergence of Cognitive Linguistics as a research paradigm, the analysis of semantic structures has been a priority on the cognitive linguistic agenda. Early studies, which shaped the field for many years, investigated the degree to which, for example, metaphor could be used to account for meaning extension. Radial categories allowed for new insights into the linguistic

[1] We would like to thank Steven Clancy, Laura Janda and the editors of this volume for a variety of useful comments on earlier drafts of this paper. The usual disclaimers apply. While working on this project Dagmar Divjak was an honorary postdoctoral fellow of the F.W.O. – Vlaanderen (Science Foundation, Flanders, Belgium); their support is gratefully acknowledged.

organization – if not also mental representation – of polysemy, and to a lesser extent near-synonymy.

Yet, despite these advances that promoted a unified analysis of semantic phenomena, semanticists working on polysemy or near-synonymy within a cognitive linguistic framework also faced problems. Polysemy requires the researcher to determine whether two usage events are identical or sufficiently similar to be considered a single sense, what the degree of similarity is between different senses, where to connect a sense to others in the network, and which sense(s) to recognize as prototypical one(s). Linguists interested in (near) synonymy do not only face these four issues, albeit at word level, but, in addition, have to decide what the differences are between the near-synonyms as well as what the relation is between semantically similar words in a domain.

By and large, three different approaches to these problems can be distinguished, i.e. (i) approaches that are not based on empirical data, (ii) approaches that are partially based on empirical data, and (iii) fully empirical approaches. We will discuss these in turn in Section 1.1, then go on to presenting the Behavioral Profile approach, our corpus-based answer to some problems of cognitive semantics (Section 1.2).

1.1. Some background

The best known example of what we, admittedly provocatively, classify as a **non-empirical** approach is Lakoff and collaborators' (1987) full-specification approach. In this approach, minimal perceived differences between usage events constitute different senses. While this approach results in very detailed lexical semantic analyses, it also leads to a difficult-to-constrain proliferation of senses and distinctions, the cognitive reality of which may be hard to establish. In an attempt to rectity some shortcomings of the full-specification approach, Kreitzer's (1997) partial-specification approach claims that minimally different usage events need not constitute different senses. However, while Kreitzer's approach is somewhat more rigorous, both approaches rely largely on decontextualized data, collected by means of introspection and analyzed using intuitions regarding what constitutes a semantically or even cognitively relevant distinction or a prototypical sense, or what the exact structure of a postulated sense network looks like.

An example of a **partially empirical** approach is Tyler and Evans' (2001) and Evans' (2005) principled-polysemy approach. This approach presents a substantial improvement over any non-empirical approach in two ways. First, the approach is more stringent since a meaning criterion rules out adding senses

ad libitum: it requires a new meaning component to be compared to other already established senses. Second, the approach makes testable predictions regarding both conceptual elaboration and grammatical distribution.

Within the class of empirical approaches, two strands can be distinguished, depending on whether the data used in the analysis are elicited or non-elicited. Prime examples of cognitive lexical semantics based on **experimental** data are Sandra and Rice (1995) as well as Rice (1996) who present lexical analyses based on data from sorting or sentence generation tasks and judgment elicitation techniques. Another example is work by Raukko (1999, 2003), who uses sentence generation and paraphrasing tasks and elicits prototypicality judgments for his polysemy-as-flexible-meaning approach (cf. Gries and Divjak, submitted, for more discussion).

Within cognitive semantics, **corpus-based approaches** that use non-elicited data are few and far between. One early cognitive-semantic corpus-based approach that is relevant for the present paper is Kishner and Gibbs' work (1996) on *just* and Gibbs and Matlock (2001) on *make*. These studies rely on collocate analysis (words at R1, i.e. the first word to the right of the head word) as well as colligation analysis (i.e., the syntactic structure of the word combination) and correlate different senses with collocations and colligations. 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). Other more recent work includes the papers published in Stefanowitsch and Gries (2006) and Gries and Stefanowitsch (2006), in particular Schönefeld (2006), who investigates the translational equivalents of the basic posture verbs *sit*, *stand*, and *lie* in English, German, and Russian with regard to how these languages have conventionalized the same physiologically determined perceptual experiences. Her investigation is based on approximately 8,000 collocations of the relevant posture verbs, and her work's crosslinguistic orientation has been a source of inspiration for the present study.

Yet, citations in corpus data have more to offer than just individual collocations and colligations: restricting the collocate/colligation analysis to the first word to the right of the head word is a heuristic that is blind to the wider syntactic structure the keyword occurs in. Corpus linguists realized this long ago, and corpus-based approaches to lexical semantics in the field of corpus linguistics have consequently produced impressive results. Take for example Atkins' (1987) study on *danger*. Her study includes an extensive collocation (words at L7 to R7, i.e. seven words to the left and seven words to the right of the head word) and colligation analysis as well as part of speech characteristics of the head word. Furthermore she introduced "ID tags", or a collocation/ colligation that correlates (probabilistically or perfectly) with a particular sense.

In his study on *urge*, Hanks (1996) extended Atkins' analysis: starting from a collocation and colligation analysis and "sense triangulation", i.e. the correlation of collocates in different clause roles, he arrives at 'Behavioral Profiles' or the totality of complementation patterns of a word that determines its semantics (Hanks 1996: 77). The full capacity of this approach is not exploited, however, and we will return to this topic in Section 1.2.

To summarize, it can be stated that, by and large, cognitive semantic studies have traditionally been based on decontextualized data, collected and analyzed by means of introspection. As a consequence, the findings may be empirically problematic: not all fine-grained sense distinctions are necessarily supported by the data (cf. Gries and Divjak, submitted). In addition, reliance on introspection as a means of data collection and intuition as the main analytic method has prevented the development of a rigorous and objectively applicable methodology. Corpus-based or computational-linguistic studies, on the other hand, have always relied on large data collections, yet their studies may be less interesting from a linguistic point of view, and this for four reasons: first, corpus-based studies are often restricted to words with few different senses or small sets of semantically similar words (*almost* vs. *nearly*, *high* vs. *tall*, *between* vs. *through*; cf., e.g, Kjellmer 2003, Taylor 2003, Kennedy 1991), they typically focus on topics that are of little interest to theoretical linguistics such as semantic prosody (cf. Xiao and McEnery 2006), they tend to be based on impoverished subsets of data available and, fourth, those data are likely noisy or skewed given (semi-) automatic preprocessing tools.

In the present study, we argue strongly for more corpus-based work in lexical semantics in general and cognitive semantics in particular, a domain that is considered by many not to be particularly well-suited for corpus-linguistic studies. We present our Behavioral Profile approach, which we believe combines the best of both the cognitive and corpus linguistic traditions, i.e. a precise, quantitative corpus-based approach that yields cognitive-linguistically relevant results. We will bring the Behavioral Profile approach to bear on polysemous near-synonyms that express 'begin' in a contrastive English-Russian analysis. Hence, the focus of the study is on presenting a corpus-based methodology that can be used to pursue cognitively-inspired lexical semantic analyses and that yields results relevant to cognitive linguistics, rather than on presenting merely a cognitive-semantic analysis of verbs that express 'begin' in itself.

1.2. Our proposal

The key assumption underlying the Behavioral Profile (BP) approach (Divjak 2003, 2004, 2006, Gries 2003, 2006, Divjak and Gries 2006, Gries and Divjak 2008, Divjak and Gries 2008, Gries and Divjak submitted) relies on the parallelism between the distributional and functional planes. Starting from the retrieval of all instances of a word's lemma from a corpus, we proceed with a (largely) manual analysis of morpho-syntactic, syntactic and semantic properties of the head word, its collocates, and the hosting clause/sentence. In other words, we extend Hanks's (1996) form of behavioral profile 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. In this way, we arrive at a maximally comprehensive behavioral profile of the lexical items studied. The BP approach differs in two important respects from previous corpus-based studies, both within and outside of cognitive linguistics: the fine granularity of the annotation goes beyond most previous work as does the subsequent evaluation of the data that is statistical in nature. The resulting semantic description is hence entirely data-driven, and delays the need for arguably personal interpretations until the very last stage of the analysis, if not bypassing it altogether.

The first purpose of the present application is to strengthen support (see the studies quoted above) for the BP approach as a powerful method that provides as *objective* a basis for the semantic analysis of both polysemous and synonymous items as possible. By *objective* we mean that even if the classification of the data points can be considered subjective to some degree,

- the amount of data to be investigated has been determined objectively, i.e. a (randomly chosen sample of a) complete concordance is used;
- the annotation of these data can be made explicit (e.g., by a set of coding instructions/criteria; cf. Table 2) and tested for consistency;
- the analysis of the data is obtained by means of statistical techniques (cf. Section 4).

This way, we postpone intuition until the stage of result interpretation (Section 5). The second purpose is to show that this approach can also be applied to the notoriously difficult area of cross-linguistic comparisons. In order to achieve this two-fold aim, the approach will be put to the test by attempting a simultaneous within-language description and across-languages comparison of polysemous and near-synonymous items belonging to different subfamilies of Indo-European, i.e. English and Russian.

2. Schmid (1993)

The current project was partially inspired by Schmid (1993), one of the first large-scale corpus-based cognitive semantic analyses. Schmid (1993) provides an in-depth corpus-based analysis of *begin* and *start* using 318 instances of the lemma *start* and 472 instances of the lemma *begin* extracted from the Lancaster-Oslo-Bergen corpus. Each instance is annotated in terms of

- the inflectional form of the verb;
- clause type: intransitive, intransitive with adverbials, transitive;
- the syntax of the complement: zero, AdvP, *ing*-clause, NP, PP, *to*-clause;
- the semantics of the subject and the semantics of the complement:[2] abstract, action, animate, cognition, human, institution, locative, manner, object, process, state, temporal;
- verb sense: 'be far from', 'begin a career', 'begin as', 'begin to speak', 'cause to begin', 'introduce', 'jump', 'protrude', 'set going in a conversion', 'set in motion', 'set out', 'set out (fig.)', 'set up', 'start a race', 'start running', 'start time unit', 'inchoative'.

Of particular interest is the correlation between the choice of phasal verb and the kind of syntactic complement. Table 1 is an excerpt of Schmid's (1993: 228) Table 4.

Table 1. Cross-tabulating transitive phasal verb with type of complement

	begin	***start***	**Totals**
NP	48	96	144
***ing*-clause**	24	53	77
***to*-clause**	256	39	295
Totals	328	188	516

These data show that *begin* correlates with *to*-constructions whereas *start* prefers *ing*-constructions (cf. Schmid 1993: 239). Following Quirk *et al.*'s (1985: 1192) classification of *to* as signalling potentiality and *ing* as indicating performance, Schmid relates those features to the respective preferences of *begin* and *start* given the strong correlation the two verbs show with *to* and *ing*

[2] Note that Schmid only distinguishes between subject and complement, but not between the semantic roles of Beginner and Beginnee. Thus, Schmid's discussion of intransitive clauses (with or without adverbials) does not distinguish agentive from non-agentive subjects.

respectively.[3] *Begin* then gives a view into the state after the onset of the action: it expresses modality/intentionality and refers to later states of affairs. It typically applies to cognitive-emotive events and non-perceivable things. *Start*, on the other hand, focuses on the actual action, the actual beginning, the very moment of transition from non-action to action. It is dynamic and applies to visible changes and actions.

[3] Quirk *et al*. (1985) and Biber *et al*. (1999) should more reliably be summarized as relating *ing*-constructions to generality, actuality, simultaneity, and direct action while relating *to*-constructions to specificity, potentiality, futurity, and indirect action. In order to be able to support/reject Schmid's claims on independent grounds, we will analyze the two dichotomies present in the dataset separately: we will first deal with the constructional complementation preferences (*ing* vs. *to*) of *begin* and *start*, then proceed to the lexical complementation preferences of *begin* and *start*. In order to arrive at a more comprehensive picture concerning the phasal verbs' preferred complementations, we decided to look at a larger corpus, the 100-million words British National Corpus World edition. With an R script, we extracted all instances of any form of *begin* or *start* tagged as a verb (either with an unambiguous tag or a portmanteau tag) followed by a word tagged as a verb gerund as well as of any form of *begin* or *start* tagged as a verb (either with an unambiguous tag or a portmanteau tag) followed by *to* (tagged as "to0") followed by a word tagged as a verb infinitive. On the basis of the concordance, we were able to look at the data in more detail in two different ways:

1) We counted how often each verb was attested with each of the two complementation patterns. The result is summarized in the table below (with expected frequencies in parentheses) and shows a highly significant correlation of *begin* with *to* and *start* with *ing* (χ^2 = 5,491.6; *df* = 1; $p <$ .001, Cramer's V = 0.38). These results conform to Schmid's (1993) LOB data as well as Wulff and Gries's (2004) ICE-GB data. Since cognitive linguistic/construction grammar approaches assume that structures co-occur (or are inserted into each other) to the extent that their meanings are compatible, this adds to the body of evidence that associates *begin* with potentiality etc. and *start* with actuality.

	begin	*start*	Totals
ing-complementation	2,780 (5,738.4)	6,239 (3,280.7)	9,019
to-complementation	21,458 (18,499.7)	7,618 (10,576.4)	29,076
Totals	24,238	13,857	38,095

2) We also counted how often each of the two phasal verbs was attested with each verb in the two complementation patterns. The resulting co-occurrence table was evaluated with a distinctive collexeme analysis using Coll.analysis 3.2 (cf. Gries 2004, Gries and Stefanowitsch 2004). The analysis by and large confirms Schmid's findings based on the semantic complement classes. *Begin* likes *be* and verbs that express cognition (*understand*, *realize*, *wonder*, *see*, *feel*, *find*, *recognize*, *suspect*, *experience*, *dawn*, *doubt*, *consider*), perception (*appear*, *emerge*, *seem*, *show*) as well as some other situations such as *fail*, *dissolve*. *Start*, on the other hand, likes less abstract verbs and takes a larger variety of verbs: it is found with basic general purpose verbs (*go*, *do*, *get*, *come*, *try*, *make*, *put*), verbs that express communicative activities (*talk*, *say*, *ask*, *cry*), and other, more dynamic activities such as *use*, *play*, *work*, *buy*, *smoke*, *look*, *fight*, *train*, *throw*, which we label 'other' because they either did not constitute a class of their own (e.g., *look* is the only verb having to do with perception) or because they are more specific than the general purpose verbs listed above but not specific enough to fall into some natural class.

Schmid's approach is comparable to our BP approach in that it involves a rather fine-grained annotation of instances obtained through an exhaustive corpus search. However, it differs in terms of the rigor with which the data are evaluated. With very few exceptions, Schmid (i) does not systematically evaluate the data statistically[4] and (ii) restricts his attention to bivariate co-occurrence patterns (sometimes even within nested tables) underutilizing the large amount of data at his disposal.

In the next section, we will demonstrate how the language-internal behavior of phasal verbs in English and Russian was investigated. In Section 4, we will show how those data can be used to contrast lemmas and senses language internally.

3. Profiling the behavior of phasal verbs

The current application of the BP approach features a contrastive analysis of 5 polysemous near-synonymous verbs that express 'begin': for English all 298 instances of the lemma *begin* and 531 of the lemma *start* were extracted from the ICE-GB; for Russian all 321 examples of the lemma *načinat'/načat'*, 173 of the lemma *načinat'sja/načat'sja*, and 156 with the lemma *stat'* were collected from the journalistic part of the Uppsala Corpus of Russian.[5]

Let us look at some examples that illustrate the main complementation patterns available: not surprisingly, and as discussed by Schmid, *begin* and *start* can be used intransitively (cf. (1a)), transitively (cf. (1b)), and as quasi-aspectual verbs (cf. (1c–d)); (2) presents the corresponding examples with *start*.

(1) a. *The land campaign has* therefore *begun*.
b. It's not thirty-two hours since *she began her shift*.

[4] Even on occasions where Schmid performs statistical tests, his methodological choices are not always optimal. For example, the interpretation of the data that are cited in connection with the most important semantic difference between *begin* and *start* (Schmid's (1993: 238) Table 6.9) involves the computation of a standard deviation for three percentages that are taken out of a larger and higher-dimensional table.

[5] We are aware of the fact that comparing written and spoken data might not be ideal, yet ICE-GB type spoken data is currently not available for Russian. In addition, the issue of how situationally-defined registers differ from each other with regard to near synonyms remains unresolved. It has been shown for English (i) that journalism resembles spoken language as far as some linguistic features, relevant in this analysis, are concerned (e.g., frequency of present perfects; cf. Gries 2006: 121) and (ii) the variation within the spoken mode and within the written mode can be so large as to make between-register variation pale by comparison (cf. Gries 2006: 121, 135).

c. And who would expect *the character to* […] *begin writing* a letter?
d. *Accidents were beginning to happen*.

(2) a. Well, *war hasn't started* yet.
b. *What* are we going to give ourselves *to start this song*?
c. I'm glad *you've started wearing* the T-shirts.
d. So *I started to think* about the Crystal Cave.

In the first two Russian examples, *načinat'/načat'* is used, once followed by an infinitive in (3) and once by a noun in (4). In (5) *načinat'sja/načat'sja* is illustrated – it only opens up a subject position – and (6) illustrates that fact that *stat'* only combines with infinitives.

(3) Буквально с первых часов космического полета *организм человека начинает приспосабливаться* к невесомости.
[Literally from the first hours of a space flight *the human organism begins to adapt* to weightlessness.]

(4) 18 апреля в Лондоне *начинает работу* общеевропейский Информационный *форум*.
[On April 18th in London the Pan-European information *forum begins its activity*.]

(5) Сафронов был лишен депутатского иммунитета, и *следствие началось*.
[Safronov was deprived of his deputy immunity and *the investigation began*.]

(6) Из двигателя общества *она стала превращаться* в тормоз его развития.
[From being the motor of society *she started to turn into* the brake of its development.]

The corpus-based method we will introduce to analyze sentences such as the ones listed above focuses on co-occurrence information of symbolic units: the symbolic unit is considered the basic unit within a cognitive linguistic approach and co-occurrences of symbolic units are easily extractable from corpora. Secondly, our approach hinges on the assumption that the words or senses investigated are part of a network of words or senses. In this network, elements which are similar to each other are connected and the strength of the connection reflects the likelihood that the elements display similar syntactic and semantic behaviour; distributional similarity is generally considered a good proxy for functional and conceptual similarity.

In total, 1,479 English and Russian sentences were annotated for 73 properties, listed summarily in Table 2; these properties capture the syntax and semantics of the verbs and their immediate surroundings.[6]

Table 2. Annotation table

Kind of ID tag	ID tag	Levels of ID tag
finite verb	lemma	*načinat'/načat', načinat'sja/načat'sja*, *stat', begin*, *start*
	aspect	Russian: imperfective vs perfective English: na
	mood	indicative, imperative, infinitive, gerund, participle, conditional/subjunctive
	tense	Russian: past, present, future English: past, present
	person	English: base form, third person singular Russian: na
	sense	have a beginning, cause to have a beginning, operate, cause to operate, first part characterized by
	voice	active, passive
complement	noun	
	verb	Russian: infinitive English: *ing*-form vs infinitive
argument structure	type	Russian: ot+gen, s+gen, s+gen (time), s+instr English: copula construction, intransitive, monotransitive, complex transitive, ditransitive, transitive
clause	type	main vs dependent
sentence	type	declarative, interrogative, exclamation
semantic roles	Beginner and Beginnee	abstract, action, animate being, change of state (self), communication, event (has natural endpoint), perception/ emotion, human being, illness, intellectual/mental, linguistic_unit (e.g., *texts*, *words*), military_action (e.g., *war*, *campaign*), motion_other, motion_other (metaphorical), motion_self, motion_self (metaphorical), social/group, perception/emotion, process (lacks natural endpoint), temporal, inanimate thing, nothing that is explicitly expressed

[6] Not all features were coded identically in both languages, yet this is solely due the nature of the corpora from which the data were retrieved and does not have any theoretical or methodological significance.

Here are some examples of the sense distinctions available for both English and Russian. Given that the senses are extremely coarse there were very few (if any) problematic cases:

(7) have a beginning: *He started accusing her.*
(8) cause to have a beginning: *She began her shift.*
(9) first part characterized by (i.e., typically the Beginnee is an activity or a time span, whose beginning exhibits characteristics that are introduced by the *begin*/*start* phrase): *One must begin with examining …*, *I had expected the day to begin with a phone call*, or *the fit itself may begin with an aura …*

In addition, English *begin* and *start* occur in the 'operate' and 'cause to operate' sense that their Russian counterparts lack. They cover a wider semantic spectrum than Russian *načinat'/načat'*, *načinat'sja/načat'sja* and *stat'*, hence are polysemous to a larger extent.

(10) operate: *It wouldn't start.*
(11) cause to operate: *He started the car.*

The result of this extensive annotation is a table with co-occurrence frequencies (for a detailed description of the procedure we refer to Gries and Divjak (forthcoming); a program for converting annotated data into behavioral-profile vectors and computing cluster-analytic statistics is now available (Gries 2008)) that contains quantitative behavioral profiles for each verb and verb-sense and can be used for quantitative analysis. In Section 4 we will present some techniques for extracting relevant semantic information from these quantitative behavioral profiles.

4. Contrasting the behavior of phasal verbs

In this section, we will rely on the information contained in the behavioral profiles for each verb and verb sense in order to arrive at an accurate and largely objective description of what defines each verb and verb sense and what differentiates them from each other. In Section 4.1, we will present contrastive

results obtained for each verb and verb sense in English; in Section 4.2, we will look at the Russian data in more detail.[7]

4.1. English

4.1.1. Plotting the BP for each lemma

Overall, there is no large difference in terms of semantic variety between *begin* and *start*: when corrected for frequency, both verbs are attested with about the same number of senses. A different picture emerges when individual ID tags are considered, however. In order to compare the individual ID tag levels of *begin* and *start*, we first computed the differences between each of the 53 behavioral profile percentages for *begin* and *start*. For example, within the ID tag 'Beginner'/'what begins', 57.82% and 37.25% of all entities doing the beginning are human beings for *start* and *begin* respectively. The difference is therefore – 20.57%, which is the largest difference between all pairwise compared percentages observed in our dataset for English and, correspondingly, reveals the most pronounced difference between *begin* and *start*. By contrast, the difference for 'Beginnee'/'what begins' for all time units (e.g., *this week*) is only 0.34%, which means that with regard to this tag *begin* and *start* behave very similarly in our data. The inspection of the differences of the behavioral profile reveals strong ID tag levels and candidates for prototypical uses:

- *begin* is preferentially used in main clauses, in the present participle/ progressive aspect, when nothing that is explicitly expressed or a concrete object (Beginner'/'what begins': thing or 0) begins to initiate a change of state of either itself ('Beginnee'/'what's begun': change of state (self)) or of something abstract ('Beginnee'/'what's begun': intellectual/mental, linguistic unit, war, abstract); other Beginnees close to the top also support that pattern: 'Beginnee'/'what's begun': event, percepts, processes, …;
- *start* on the other hand is preferentially used transitively and in subordinate clauses, in the infinitive, when a human instigator ('Beginner'/'what begins': human) causes an action (in particular communicative actions) to take place or, a bit further down the list, causes a concrete object to operate (which nicely corresponds to the specific sense of *start* exemplified in *He started the bike*).

[7] All computations and graphs were performed and created with R for Windows 2.5.1 (cf. R Development Core Team 2007).

In order to summarize the overall tendencies in a way that allows for a straightforward comparison with the Russian data in the following section, Figure 1 summarizes the behavioral profile for both *begin* and *start* using only the ID tags that apply to both the English and the Russian data. Each ID tag and its level (separated by three underscores) is plotted on the *y*-axis coordinate of the difference between that ID tag's value for *begin* and that ID tag's value for *start*. The ID tags are rank-ordered on the basis of the size of that difference from left to right with the ranks displayed on the *x*-axis. Highlighting (in right or left) indicates the most distinctive ID tags for each verb. For example, the rightmost data point at –0.21 indicates that the entity *beginning* an action is a human 37% of the time, while the entity *starting* an action is a human 58% of the time, hence there is a difference of 21%.

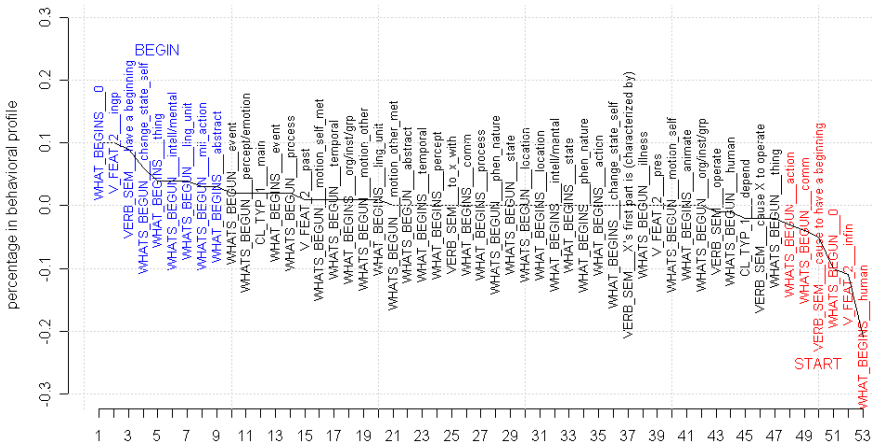

Fig. 1. Snake plot of the most distinctive ID tags for each verb

In the next section, we will briefly address the issue of how different senses of *begin* and *start* are related to each other.

4.1.2. Clustering the BPs for each sense

Contrary to Schmid, and in line with more recent and more rigorous attempts to avoid positing large numbers of senses, we have restricted our sense annotation to the five high-level senses listed in Table 2. As in our previous work, we have applied a hierarchical agglomerative cluster analysis to the complete behavioral profile (similarity measure: Canberra metric, amalgamation

rule: Ward's algorithm); we found that the five different verb senses ('have a beginning', 'cause to have a beginning', 'first part characterized by', 'operate', 'cause to operate') form two clear clusters. The left panel of Figure 2 contains the dendrogram showing these two clusters, while the right panel visualizes average and individual silhouette widths (the grey step function lines and the vertical black lines respectively) for all possible cluster solutions;[8] the maximum silhouette width is obtained for a two cluster solution, as the numbers in the plot specify.

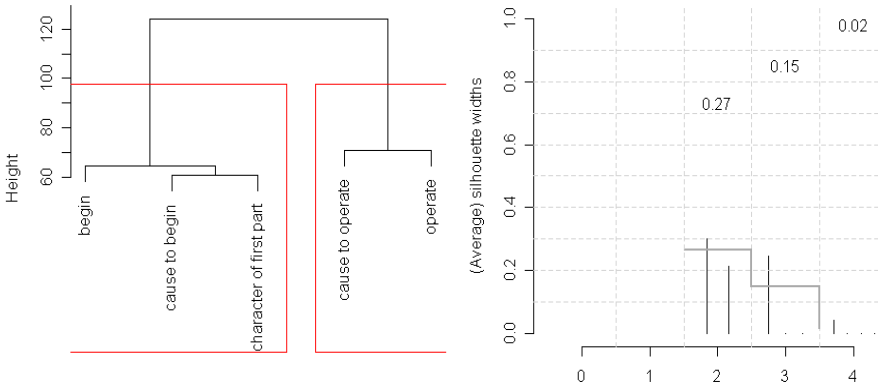

Fig. 2. Cluster-analytic results of begin's and start's senses

Here, different (high-level) senses of the verbs *begin* and *start* are clustered on the basis of overall semantic similarity, while in Gries (2006) different fine-grained senses of the verb *run* were mainly distinguished in terms of causativity/transitivity. This result may in large part be due to the fact that the five senses of the two phasal verbs here exhibit a very high degree of semantic similarity – much more so than many of the semantically very different senses of the verb *run*. When overall semantic similarity is low, distinguishing senses along a fundamental parameter such as causativity/transitivity can account for much of the variation. However, when overall semantic similarity is high, as with the phasal verbs, causativity/transitivity simply does not add much discriminatory power – lower-level semantic features are more important.

[8] Silhouette widths are a means to assess the quality of a cluster-analytic solution. They are based on comparing the average similarity of an element to the other elements in the same cluster to the average similarity of an element to the elements in all other clusters. The larger the average silhouette width for a solution, the better the clustering; cf. Kaufman and Rousseeuw (1990) for details.

In the following section, we will perform similar analyses for our Russian data and finally, proceed to comparing the results cross-linguistically in Section 5.

4.2. Russian

Given that Russian has three verbs to express 'begin', this section will be structured slightly differently. First, we will cluster the three verbs *načinat'/načat'*, *stat'* and *načinat'sja/načat'sja* to find out which verbs are more similar to each other (Section 4.2.1); next (Section 4.2.2), we will plot the BPs of the three verbs against each other to reveal in which respects these verbs resemble each other and in which respects they differ from each other. Finally, we will turn to analyzing the verbs' senses (Section 4.2.3): clustering the BPs facilitates selecting the most distinctive ID tags per verb sense.

4.2.1. Clustering verbs

Out of all 73 assignable ID tag levels, 69 are contained in the behavioral profile (BP) for *načinat'/načat'*, 41 in the BP for *načinat'sja/načat'sja* and 44 in the BP for *stat'*. Yet, given that *načinat'/načat'* is by far the most frequent, i.e. it is used in 50% of all occurrences, all three verbs should be considered equally versatile.

Cluster analysis shows that *načinat'/načat'* and *stat'* form one cluster, while *načinat'sja/načat'sja* is distinct. This partition reveals a strong influence of the semantics of the argument structure on the clustering, i.e. *načinat'sja/načat'sja* lacks a syntactic/semantic subject position, i.e. it lacks a Beginner, whereas *načinat'/načat'* and *stat'* have both Beginner and Beginnee. We will return to this difference below.

4.2.2. Plotting the verbs' BPs

As above for *begin* and *start*, we computed pairwise differences of behavioral profile percentages of ID tag levels. In order to be able to compare the English and the Russian data, we provide a snake plot of the ID tags attested in both languages (hence aspect will not show up in the plot). Figure 3 contrasts *načinat'/načat'* and *stat'*.

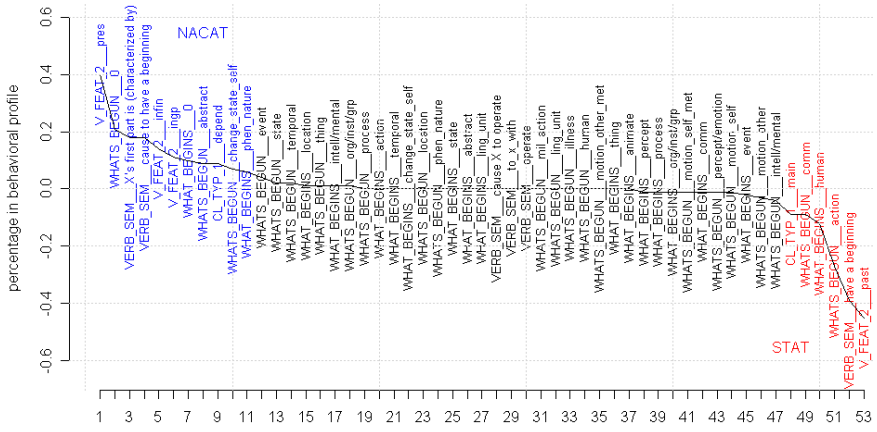

Fig. 3. Snake plot of the most distinctive ID tags for načinat'/načat' vs stat'

The following is a discussion of the top 17 distinctive properties between the behavioral profiles for the two most similar verbs *nacat'* and *stat'*.

- *Načinat'/načat'* differs from *stat'* in that it is found in the imperfective, as a gerund or infinitive, and in the present tense; it is often found in combination with *s* followed by a genitive ('since'), expressing a situation that has a clear source or begins at a specific moment in time. The beginning applies to both nouns and verbs expressing abstract concepts and changes of state instigated by the unknown or by nature. *Načinat'/načat'* expresses all three senses, i.e. "have a beginning", "the first part is characterized by" as well as something has been "caused to have a beginning".
- *Stat'*, on the other hand, prefers the perfective, the indicative and the past, is instigated by human beings and is aimed at actions in general as well as at communicative activities. Different from *načinat'/načat'*, *stat'* is restricted to expressing that something, in particular an event, has a beginning. Supporting this finding is that fact that *stat'* is never encountered with nouns or without Beginnee altogether.

Several of the differences revealed by an analysis of the BP correspond to traditional interpretations (Flank 1987, Paillard 1998, Dickey 2000, Padučeva 2001) claiming that the verbs differ with respect to the phase of action that is referred to: *stat'* is said to defocus the beginning and to express a smooth transition into a new state, whereas *načinat'/načat'* is typically thought to foreground the beginning as an independent event.

Nacat' expresses a situation that begins at a specific moment in time; the beginning it expresses is ongoing and can be observed, hence it is an action in its

own right. The preferences of *stat'*, on the other hand, can be interpreted as an indication of the fact that *stat'* itself expresses a completed action. At the same time, *stat'* is restricted to expressing that something has a beginning: it requires a second action in order to render phasal meaning (in combination with a noun in the instrumental case, *stat'* means "become"). The reliance of phasal *stat'* on that second action backgrounds the beginning while foregrounding the second event.

Načinat'/načat' differs from the third verb to which it is morphologically related, *načinat'sja/načat'sja.* The properties that distinguish best between *načinat'/načat'* and *načinat'sja/načat'sja* are summarized in Figure 4.

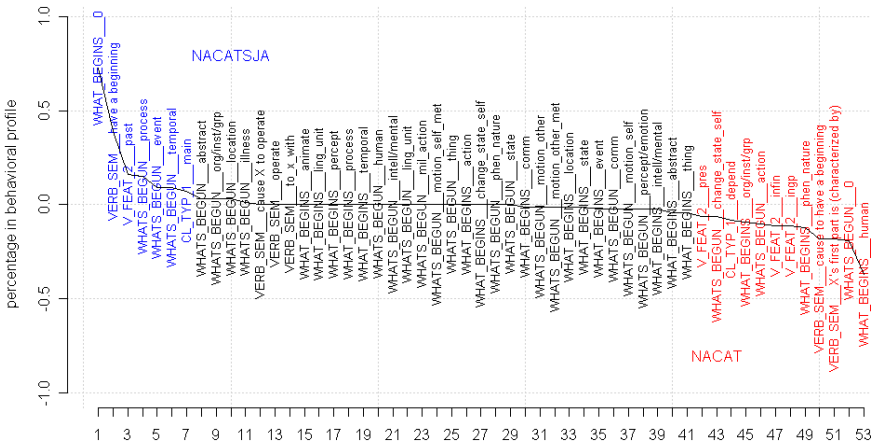

Fig. 4. Snake plot of the most distinctive ID tags for načinat'sja/načat'sja vs načinat'/načat'

- the first distinctive property for *načinat'sja/načat'sja* is the zero Beginner, the second distinctive property signals that *načinat'sja/načat'sja* is restricted to expressing the meaning "have a beginning"; the following properties highlight that this meaning is typically expressed in a main clause, with *načinat'sja/načat'sja* in the past tense; *načinat'sja/načat'sja* applies to processes, events and time-related situations.
- compared to *načinat'sja/načat'sja*, *načinat'/načat'* favors the present tense and dependent clauses; it conveys a wider range of senses including "causing something to have a beginning" and "characterizing the first part of X"; *načinat'/načat'* is instigated by human beings, groups/institutions or phenomena of nature and is typically applied to actions and changes of state that affect the self, yet the Beginnee does not need to be explicitly expressed.

Načinat'sja/načat'sja, being restricted to a Beginnee, lacks an overt active Beginner. These argument structure restrictions signal that *načinat'sja/načat'sja* embodies an externally imposed, agentless beginning (cf. Padučeva 2001: 34); it applies exclusively to obligatory nominal Beginnees that express events and processes. *Načinat'/načat'*is *načinat'sja/načat'sja'*s inverse: it has slots for both Beginner and Beginnee, and can choose to fill up either or both positions with elements from a variety of semantic groups.

In this section, we have briefly touched upon how the three verbs, i.e. *načinat'/načat', načinat'sja/načat'sja* and *stat'*, correlate with the three different senses 'have a beginning', 'cause to have a beginning' and 'characterize the first part of X'. In the next section, we will investigate how the senses themselves relate to each other and which ID tags and levels of ID tags are involved in defining and distinguishing senses.

4.2.3. Clustering verb senses

In Russian, all three verbs express the sense 'have a beginning', whereas only *načinat'/načat'* can render 'cause to have a beginning' and 'characterize the first part of X'. Yet, a hierarchical agglomerative cluster analysis (dendrogram not shown) reveals that the senses 'have a beginning' and 'cause to have a beginning' cluster together – as they do for English – and are more similar to each other than either of them is to the sense 'characterizing the first part of X'. Clearly, argument structure plays an important role here and semantic differences caused by diathesis alternations are rightly considered highly influential.

Grammatically speaking, in Russian, the sense 'have a beginning' scores high (values ranging from 0.6 up to 0.99) for the following ID tags and levels: declarative main clause, perfective indicative finite verb and active infinitive. 'Cause to have a beginning' is typically expressed by perfective *načat'* when combined with a noun and is encountered in the indicative mood in declarative main clauses. Finally, the sense 'characterizing the first part of X' is most typical of declarative sentences with imperfective present *načinat'* followed by a noun.

As far as the semantic roles of Beginner and Beginnee are concerned for these three Russian verbs, human beings, phenomena of nature and nothing that is explicitly expressed tend to begin actions, processes and mental activities. Human beings, social entities or nothing that is explicitly expressed cause a process or event expressed by a noun or a communicative act to have a beginning. And finally, human beings and nothing that is explicitly expressed tend to have a first part that is characterized by or begun with or from something.

5. A cognitive cross-linguistic comparison

Overall, the within-language similarity between verbs is higher than between language similarity: in an across language cluster analysis including only properties attested in both languages *begin* and *start* cluster together as *načinat'/načat'* and *stat'* do, while *načinat'sja/načat'sja* is kept separately.

Yet, the underlying dissimilarity matrix reveals that while *begin* and *start* may well be most similar to each other within one language, seen across languages *načinat'/načat'* and *begin* are most similar. From this it does, however, not follow that *stat'* and *start* are equivalent. When we look at *stat'* in more detail, we see that it does indeed resemble *start* in that it prefers the past tense and certain types of Beginnee, i.e. actions, communications and mental activities. At the same time, *stat'* resembles *begin* in that it highlights the "view into the state after the onset of the action", as we characterized *begin*. These findings highlight the fact that a one-to-one lexical and/or conceptual mapping between phasal verbs in Russian and English may well be absent, a conclusion that is supported by our overall conceptual findings.

The prototype for each verb and set of verbs in each language seems to revolve around a different set of characteristics altogether: this difference is clearly revealed by the snake plots. In English (Figure 1), 12 out of the 15 most distinctive properties for *begin* and *start* relate to the type of Beginner and Beginnee; English *begin* is concerned with more abstract, less tangible/non-perceivable processes whereas *start* is associated with more dynamic and concrete actions instigated by humans. In Russian (Figures 3 and 4), however, only 6 out of the 17 (for *načinat'/načat'* versus *stat'*) and 8 out of the 19 (for *načinat'/načat'* versus *načinat'sja/načat'sja*) most distinctive ID tags relate to lexical preferences of the phasal verbs; the majority of distinctive properties relates to the aspectual and argument structure peculiarities of the verbs. The two Russian verbs *načinat'/načat'* and *stat'* differ with respect to the phase of action that is referred to: given that each of the structural verb-related differences account for a relatively large portion of the variation between *stat'* and *načinat'/načat'*, *stat'* can be said to defocus the beginning and to express a smooth transition into a new state, whereas *načinat'/načat'* is typically thought to foreground the beginning as an independent event. The difference between *načinat'/načat'* and *načinat'sja/načat'sja*, on the other hand, clearly revolves around the concept's compatibility with agentivity: *načinat'/načat'* does take Beginners, *načinat'sja/načat'sja* does not. In other words, the difference between *begin* and *start* may be termed lexical, whereas that between *načinat'/načat'* and *stat'* seems aspectual in nature (cf. Zaliznjak & Šmelev 2002: 219–222, and in

less overt terms Černova 1999: 165–169), and the difference between *načinat'/ načat'* and *načinat'sja/načat'sja* relates to argument structure.

Verbs that express 'begin' in English and Russian are but one example of languages' tendency to carve up (a similar) conceptual space in a unique way and to opt for a different division of labour between the lexemes available to express similar concepts. This can lead to dissimilarities that are of an entirely different order and are bound to be overlooked by comparative cognitive semanticists unless a methodology is used that adequately captures the multivariate nature of the phenomenon; behavioral profiling does exactly that.

6. Conclusion

In this study, we have put our BP approach to the test by applying it to the study of polysemous near synonyms in English and Russian that express 'begin'. We hope to have achieved three objectives.

First of all, we have provided a detailed usage-based characterization of *begin*, *start*, *načinat'/načat'*, *načinat'sja/načat'sja* and *stat'* that captures what we have termed the verbs' behavioral profile. Snake plot summaries facilitate an immediate identification of the most central usage characteristics of these five verbs and of the related prototypical scenarios they evoke. These findings are of interest to both theoretically-oriented cognitive linguists and practically-oriented lexicographers. Secondly we have illustrated how, within a semantically homogenous set of verbs and senses, clusters of verb senses exist, and these may be revealed on the basis of distributional characteristics collected in BPs. The skeptic might argue that a cluster analysis will *always* yield clusters. While this is true, there are many possible cluster solutions, but only a few make sense, and the ones we obtained are probably the most or the second-most sensible solutions one could have wished for. Finally, we have shown how the objective annotation of comparable semantic and distributional ID tags of translational equivalents across languages enables us to take first steps toward a more rigorous cross-linguistic and cognitive-linguistic analysis.

While corpus-linguistic methods are more frequently applied in cognitive linguistics than in many other linguistic frameworks, their utility has not been sufficiently, let alone uniformly, recognized. Among others, Raukko (1999, 2003) launches a ferocious attack at corpus linguistic methods, but such criticism typically throws out the baby with the bathwater especially now that corpus linguistics is evolving towards a methodologically more mature and quantitatively more sophisticated state; the present study applies some of these

quantitatively more refined methods to the study of near synonyms in a cognitive-linguistic framework.

Quantitative rigor does not imply that the job is done, however. There are both methodological and conceptual steps left to take. We have illustrated, albeit in an early footnote, that both the BP approach to *begin* and *start* as well as previous approaches are supported and enriched by additional analyses. The analysis of data from a different and much larger corpus than has been used for this paper and with a method that has been shown to be sensitive to lexical and constructional semantics confirms the patterns and interpretations found as well as the conclusions derived from the behavioral profiles. On the methodological side, the results from behavioral profiles must be tested against other empirical data, and the level of detail present in an objective data collection of this kind makes this task feasible. In the present case, for example, additional corpus data from English-Russian parallel corpora may help to determine how well our BP approach has succeeded in identifying the relevant dimensions of variation and hence in predicting the choice for one or the other available alternative. Alternatively, a reasonable next step would be to pursue a more fine-grained sense coding (possibly along the lines of Schmid 1993) to see with which distributional patterns, if any, the sense distinctions are correlated. The data-driven nature and comprehensiveness of behavioral profiles offer many different avenues of research, waiting to be explored by cognitive linguists.

REFERENCES

Atkins, Beryl T. Sue (1987). "Semantic ID tags: Corpus evidence for dictionary senses". *Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary*: 17–36.

The British National Corpus, version 2 (BNC World) (2001). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk

Černova, S. V. (1999). "Fazisnye glagoly kak predikaty otražajuščie celenapravlennuju dejatel'nost' čeloveka". *Semantika. Funkcionirovanie. Tekst.* Kirov: 158–173.

Dickey, S. (2000). *Parameters of Slavic Aspect: A Cognitive Approach*. Stanford: CSLI Publications.

Divjak, D. (2003). "Ways of Planning Actions. A Cognitive Semantic Approach to Near Synonyms in Russian". Paper presented at the International Cognitive Linguistics Conference 2003. Universidad de La Rioja, 22 July 2003.

Divjak, D. (2004). *Degrees of verb integration. Conceptualizing and categorizing events in Russian*. Unpublished Ph.D. Dissertation, K.U. Leuven (Belgium).

Divjak, D. (2006). "Ways of intending: Delineating and structuring near synonyms". In: Gries, S. Th. & A. Stefanowitsch (eds.). *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*. Berlin, New York: Mouton de Gruyter: 19–56.

Divjak, D. & S. Th. Gries (2006). "Ways of trying in Russian: clustering behavioral profiles". *Corpus Linguistics and Linguistic Theory* 2(1): 23–60.

Divjak, D. & S. Th. Gries (2008). "Clusters in the mind? Converging evidence from near-synonymy in Russian". *The Mental Lexicon* 3(2): 188–213.

Evans, Vyvyan (2005). "The meaning of *time*: polysemy, the lexicon and conceptual structure". *Journal of Linguistics* 41(1): 33–75.

Flank, Sharon (1987). "Phase subdivisions and Russian inceptives". *Die Welt der Slaven* 32(2): 310–16.

Gibbs, Raymond W. Jr. & Teenie Matlock (2001). "Psycholinguistic perspectives on polysemy". In: Cuyckens, H. & B. Zawada (eds.). *Polysemy in Cognitive Linguistics*. Amsterdam, Philadelphia: John Benjamins: 213–39.

Gries, S. Th. (2003). "The many meanings of *to run*: cognitive linguistics meets corpus-based linguistics". Paper presented at the International Cognitive Linguistics Conference 2003. Universidad de La Rioja, 22 July 2003.

Gries, S. Th. (2004). Coll.analysis 3.2. A program for R for Windows.

Gries, S. Th. (2006). "Corpus-based methods and cognitive semantics: The many meanings of *to run*". In: Gries, S. Th. & A. Stefanowitsch (eds.). *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*. Berlin, New York: Mouton de Gruyter: 57–99.

Gries, S. Th. (2006). "Exploring variability with and between corpora: some methodological considerations". *Corpora* 1(2): 109–51.

Gries, S. Th. (2008). BP. A program for R (for Windows).

Gries, S. Th. & D. Divjak (forthcoming). "Behavioral profiles: a corpus-based approach to cognitive semantic analysis". In: Evans, Vyvyan & Stephanie S. Pourcel (eds.). *New directions in cognitive linguistics*. Amsterdam, Philadelphia: John Benjamins.

Gries, Stefan Th. & Dagmar Divjak (submitted). "Quantitative approaches in usage-based cognitive semantics: myths, erroneous assumptions, and a proposal".

Gries, Stefan Th. & Anatol Stefanowitsch (2004). "Extending collostructional analysis: a corpus-based perspective on 'alternations' ". *International Journal of Corpus Linguistics* 9(1): 97–129.

Gries, S. Th. & A. Stefanowitsch (eds.) (2006). *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*. Berlin, New York: Mouton de Gruyter.

Hanks, P. (1996). "Contextual dependency and lexical sets". *International Journal of Corpus Linguistics* 1(1): 75–98.

Janda, L. A. (to appear). "What is the role of semantic maps in cognitive linguistics?" In: P. Stalmaszczyk and W. Oleksy (eds.). *Festschrift for Barbara Lewandowska-Tomaszczyk*.

Kennedy, Graeme (1991). "*Between* and *through*: the company they keep and the functions they serve". In: Aijmer, K. & B. Altenberg (eds.). *English Corpus Linguistics*. London: Longman: 95–110.

Kishner, J. M. & Raymond W. Gibbs Jr (1996). "How *just* gets its meanings: Polysemy and context in psychological semantics". *Language and Speech* 39(1): 19–36.

Kjellmer, G. (2003). "Symomy and corpus work: on *almost* and *nearly*". *ICAME Journal* 27:19–27.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. Chicago, IL: The University of Chicago Press.

Kreitzer, A. (1997). "Multiple levels of schematization: a study in the conceptualization of space". *Cognitive Linguistics* 8(4): 291–325.

Norvig, P. & G. Lakoff (1987). "Taking: a study in lexical network theory". In: Aske, J., N. Beery, L. Michaelis, & H. Filip (eds.). *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*. Berkeley, CA: BLS: 195–206.

Padučeva, E. V. (2001). "Fazovye glagoly i semantika načinatel'nosti". *Izvestija Akademii Nauk. Serija literatury i jazyka* 60 (4): 29–39.

Paillard, D. & F. Fici Giusti (1998). "L'inchoation en Russe: entre préverbes et auxiliaries". *Le Langage et l'Homme* 33(1): 79–94.

Quirk, R., S. Greenbaum, G. Leech, & J. Svartvik (1985). *A Comprehensive Grammar of the English Language*. London, New York: Longman.

R Development Core Team (2006), *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Raukko, J. (1999). "An 'intersubjective' method for cognitive-semantic research on polysemy: The case of *get*". In: Hiraga, Masako K., Chris Sinha, & Sherman Wilcox (eds.). *Cultural, Psychological and Typological Issues in Cognitive Linguistics*. Amsterdam, Philadelphia: John Benjamins: 87–105.

Raukko, J. (2003). "Polysemy as flexible meaning: Experiments with English get and Finnish pitää". In: Nerlich, B., Zazie T., Vimala H. & David D. Clarke (eds.). *Polysemy: Flexible Patterns of Meaning in Mind and Language*. Berlin: Mouton de Gruyter: 161–93.

Rice, S. (1996). "Prepositional prototypes". In: Pütz, M. & R. Dirven (eds.). *The Construal of Space in Language and Thought*. Berlin, New York: Mouton de Gruyter: 135–65.

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.

Schmid, H.-J. (1993). *Cottage and co., idea, start vs. begin*. Tübingen: Max Niemeyer.

Schönefeld, D. (2006). "From conceptualization to linguistic expression: where languages diversify". In: Gries, S. Th. & A. Stefanowitsch (eds.). *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*. Berlin, New York: Mouton de Gruyter: 297–344.

Stefanowitsch, A. & S. Th. Gries (2006). *Corpus-based approaches to metaphor and metonymy*. Berlin, New York: Mouton de Gruyter.

Taylor, J. R. (2003). "Near synonyms as co-extensive categories: *high* and *tall* revisited". *Language Sciences* 25(3): 263–84.

Tyler, A. & V. Evans (2001). "Reconsidering prepositional polysemy networks: The case of *over*". *Language* 77(4): 724–65.

Xiao, R. & A. McEnery (2006). "Collocation, semantic prosody and near synonymy: a cross-linguistic perspective". *Applied Linguistics* 27(1): 103–29.

Wulff, S. & S. Th. Gries (2004). "*Prefer to construe* vs. *prefer construing*: a corpus-linguistic perspective on non-finite sentential complementation". Paper presented at Current Trends in Cognitive Linguistics. University of Hamburg, 11. December 2004.

Zaliznjak, A. A. & Šmelev, A. D. (2002). „Semantika 'načala' c aspektologičeskoj točki zrenija". In: *Logičeskij analiz jazyka. Semantika načala i konca.* Moskva: Indrik: 211–224.