Clusters in the mind?

Converging evidence from near synonymy in Russian*

Dagmar Divjak and Stefan Th. Gries
University of Sheffield (UK), F.W.O. Vlaanderen (Belgium) / University of California, Santa Barbara (USA)

This paper provides experimental evidence to support the existence of mental correlates of lexical clusters. Data were collected by means of a sorting task and a gap filling task designed to study the cognitive reality of clusters of near synonyms as well as of the properties that have high predictive power for subcategorizing near synonyms. The results for nine near-synonymous verbs expressing 'try' in Russian confirm the linguistic account of the synonym structure that was proposed on the basis of corpus data in Divjak and Gries (2006). We conclude that speakers learn and retain exemplars from which they extract distributional patterns that help shape the arrangement of verbs in lexical space. Consequently, a corpus-based behavioral profile approach to lexical semantics is strengthened as it provides a firm basis for cognitively realistic language descriptions.

Keywords: near synonymy, behavioral profiles, cluster analysis, lexical clusters, mental correlates of linguistic models, Russian, sorting, gap-filling

> The notion of priming as here outlined assumes that the mind has a mental concordance of every word it has encountered, a concordance that has been richly glossed for social, physical, discoursal, generic and interpersonal context. This mental concordance is accessible and can be processed in much the same way that a computer concordance is, so that all kinds of patterns, including collocational patterns, are available for use. It simultaneously serves as a part, at least, of our knowledge base (Hoey, 2005, p. 11).

The Mental Lexicon 3:2 (2008), **188–213**. DOI 10.1075/ml.3.2.03div
ISSN 1871–1340 / E-ISSN 1871–1375 © John Benjamins Publishing Company

Cognitively real(istic) linguistics

One of the areas that facilitated the emergence of cognitive linguistics as a new research paradigm was that of lexical semantics. Cognitive linguists strive to make their "account of human language accord with what is generally known about the mind and the brain, from other disciplines as well as our own" (cf. Lakoff's Cognitive Commitment, Lakoff, 1990, p. 53). Hence, early lexical semantic studies, which shaped the field for years to come, investigated the degree to which, for example, metaphor could be used to account for meaning extension; similarly, the concept of radial categories allowed for new insights into the linguistic organization and related mental representation of polysemy, and to a lesser extent near synonymy. This approach increased the expectation, yet not necessarily the likelihood, of being able to find mental correlates for linguistic models. Although the field witnessed a gradual shift from intuition-based, corpus-illustrated work to corpus-based analyses (cf. Gibbs & Matlock, 2001; Kishner & Gibbs, 1996 and the papers in Gries & Stefanowitsch, 2006 and Stefanowitsch & Gries, 2006), few lexical semanticists have taken on the challenge of validating relevant linguistic findings experimentally (but see Arppe & Järvikivi, 2007; Rice, 1996; Sandra & Rice, 1995).

Some of the above publications criticized cognitive linguistic methodology, and in particular the widely used network representations of words and word senses, for a number of reasons. Among the most pressing questions clearly are: which elements of usage need to be captured to arrive at an objective and satisfactory description of meaning? And what, if any, contribution can linguistic work on polysemy or near synonymy make to issues of mental representation of lexical items? This paper seeks to remedy these issues by constructing a model for nine Russian near synonyms expressing 'try' that is based on corpus data and is validated experimentally.

A corpus-based approach to meaning

In recent work, the "behavioral profile"-approach, henceforth BP approach, to lexical semantics was introduced (see Gries and Divjak, in press, for an overview). Our principal method of investigation extracts every clue possible (which we, following Atkins, 1987, refer to as "ID tags") from the corpus sentences in which the verbs under consideration are used so as to infer different facets of their meanings and uses. In this particular case, these ID tags comprise formal characteristics of the verb and the clause or the sentence the finite verb occurs in, elements that co-occur with the verb (such as adverbs, particles and connectors) as well as paraphrases (i.e., characterizations) of the semantic properties of the subject and infinitive (see section *Tagging for meaning*). Taken together, these ID tags form what we, modifying a term coined by Hanks (1996, p. 79), refer to as the "behavioral profile" for each verb. In other words, the BP approach takes a usage-based view of meaning, and therefore we will use the words *use* and *meaning* interchangeably. While differences in usage can be syntactic, semantic, pragmatic or socio-lectal in nature, we will restrict our discussion to denotational aspects of meaning, thus leaving aside pragmatic and socio-lectal variation.

Since the BP approach is usage-based, it qualifies as a data-driven and hence more objective means to capture and compare the meanings (Divjak, 2006; Divjak & Gries, 2006) or senses (Gries, 2006) of words. In addition, behavioral profiles can be subjected to exploratory statistical techniques such as cluster analyses that help to uncover internal structure in large datasets.

Tagging for meaning

Divjak and Gries (2006) analyzed 1,585 sentences, each containing one of nine verbs that, in combination with an infinitive, express 'try' in Russian: *po/probovat'* ('try'), *pytat'sja* ('try, attempt'), *starat'sja* ('try, endeavor'), *silit'sja* ('try, make efforts'), *norovit'* ('try, strive to, aim at'), *poryvat'sja* ('try, endeavor'), *tščit'sja* ('try, endeavor'), *pyžit'sja* ('go all out') and *tužit'sja* ('make an effort, exert oneself'). All 1,585 examples (between 100 and 250 per verb, depending on the frequency of the verb) were annotated for 87 properties, a.k.a. levels of ID tags. The ID tags are listed in Table 1, together with their most frequently encountered levels.

Coding started from observable formal characteristics of the finite verb and was gradually extended to include information on other elements of the sentence. In a first coding round, we zoomed in on the elements present in all constructions built on the [V_{fin} V_{inf}] pattern, i.e. the aspect, mode and tense of the finite verb. In addition, we coded those elements that are strictly necessary to form a full-fledged sentence, i.e. information on the type of clause the [V_{fin} V_{inf}] sequence is used in and, linked to the main- or subclause status of the sentence the finite verb occurs in, the case marked on the subject slot. Taken together, the structural data on clause type and related form of the subject, as well as details on the aspect, mode and tense of the verbs in the [V_{fin} V_{inf}] sequence, form the skeleton of the sentence. From here, one can fill up constructional slots with lexical elements.

In the second round, the adverbs, particles and connectors that are used in the corpus sample are at the center of attention. Detecting adverbs, particles and connectors does not require semantic intuitions, but is nevertheless semantically informative. For example, verbs combine with a whole range of adverbs, particles or connectors, but not all verbs prefer identical (sets of) adverbs; in addition, adverbs

Table 1. ID tags used in annotating corpus extractions (adapted from Divjak & Gries, 2006)

Type 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
syntactic	sentence type	declarative, exclamative, imperative, interrogative
	clause type	main vs. dependent
semantic	semantic type of subjects	concrete vs. abstract, animate (human, animal) vs. inanimate (event, phenomenon of nature, body part, organization/institution, speech/text) etc.
	properties of the process denoted by the verb	physical actions, perception, communication, intellectual activities, emotions, wishes/desires etc.
	controllability of infinitive action	high vs. medium vs. no controllability
	adverbs, particles, connectors	temporal, locative, etc.
	negation	present vs. absent, attached to which element

and particles often provide information that characterizes a particular situation or action.

The third type of information that has been coded is the most semantic in nature. It contains semantic paraphrases for the subject and the infinitive, typical candidates for a traditional collocation analysis. Within the scope of this study, we systematically classified the nominative subject paradigms along a combination of lines, i.e. the opposition animate vs. non-animate including insects and the distinction between addressable, i.e. human, and non-addressable or animal animate subjects. There are some additional distinctions for non-animates, i.e. concrete vs. abstract and further specifications have been made on the basis of the kinds of subjects used in the data sample, i.e. man-made and non-man-made concrete things, the latter being most often phenomena of nature (e.g. *the sun, the earth*) or body parts, as well as abstract concepts (e.g. *an idea, an insight*) and groups or organizations. For the infinitives, we adopted a labeling system that is inspired by the "semantic primitives of human behavior" set forth in Apresjan's (1995) linguistic naïve world view. The eight "basic systems of a human being" that Apresjan (1995, p. 355–6) distinguishes are comparable to "basic domains" (Langacker, 1987, Chapter 4) or the semantic primitives underlying work on the Natural Semantic Metalanguage by Wierzbicka and her collaborators, i.e., domains that are not characterized in terms of other more fundamental domains (cf. e.g., Wierzbicka, 1996). These semantic primitives of human behavior are generalizations over the paraphrasing semantic labels for classifying infinitives we use in our research. Although strictly speaking many other properties could be tagged for, they have not been included in the analysis as either some of these other tags are already included indirectly, or they do not apply in this particular case of verbs, or do not lend themselves well to operational definitions.

The overall distributional behavior of the nine verbs was summarized in a table of co-occurrence frequencies. Put differently, each verb's distribution is characterized by a vector of percentages that represents how often a particular verb co-occurs with each of the levels of the ID tags listed above. This dataset was subjected to a hierarchical agglomerative cluster analysis, using the Canberra similarity metric and Ward's amalgamation strategy (for a more precise description of the procedure, cf. Gries & Divjak, in press). In the resulting dendrogram presented in Figure 1, verbs that are clustered or amalgamated early are semantically similar, whereas verbs that are amalgamated late are rather dissimilar. For example, *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. The dendrogram also gives an indication of how autonomous the groups of verbs are: the distance between different points of amalgamation is a function of the independence of a given lower-order cluster from the next higher-order cluster. In the present case, the plot clearly consists of three clusters; on the basis of the ID tag levels that were most strongly correlated with these clusters, Divjak and Gries (2006) labelled the clusters YOU COULD SUCCEED, YOU WON'T SUCCEED and YOU CAN´T SUCCEED, based on the likelihood of success.

BPs are not only an excellent basis for revealing the internal structure of a group of near synonyms in a way consistent with fundamental cognitive linguistic assumptions. What is more, they allow us to investigate the nature of the three categories suggested by the dendrogram more thoroughly. Between-cluster similarities and differences were inspected using *t*-values that pick out those variables that discriminate well between clusters, i.e., they foreground the most important properties of a cluster.[1] More specifically, *t*-values facilitate determining which variables are most strongly represented (in the case of high positive *t*-values) and which variables are most strongly underrepresented (in the case of low negative *t*-values) in a particular cluster. The higher the *t*-value for a certain property in a particular cluster, the higher the chance that a verb from this cluster will be used to denote this property. In the following section, we discuss the top 25 most revealing scores, i.e. the variables having positive *t*-values for one cluster and negative *t*-values for the other two clusters and vice versa; cf. Divjak and Gries (2006) for details.

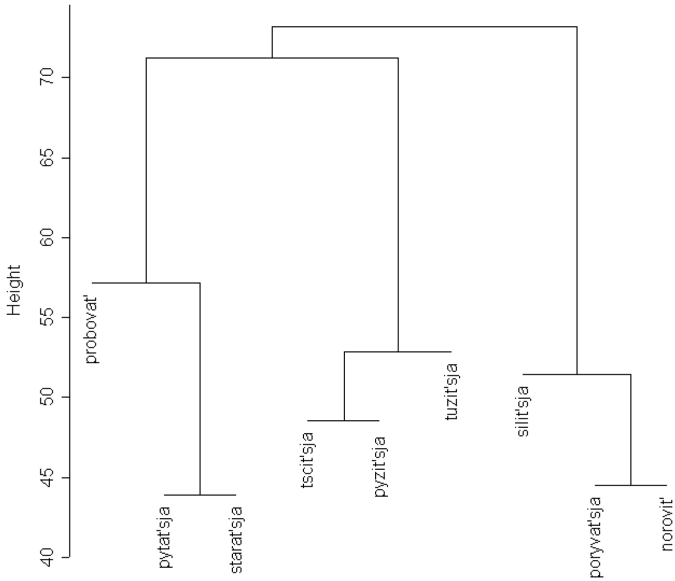

Figure 1. Dendrogram of nine Russian verbs meaning 'try' (from Divjak and Gries 2006).

Evaluating the results

If we pull together the dimensions with the most revealing *t*-values for the arguably most central and neutral, i.e. most widely applicable, cluster and incorporate them into one scenario, the characterization that emerges for *pytat'sja*, *starat'sja* and *probovat'* is the following:[2] a human (rather than an animal or an insect) is exhorted to undertake an attempt to move himself or others (rather than to undertake mental activities); often, these activities are negated. All three verbs 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 attracted optional collocates express that the subject got permission to carry out the infinitive action (using *pust'*, $t = 1.008$), that the attempt was brought to an untimely 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 cluster with *silit'sja*, *poryvat'sja* and *norovit'*, an inanimate subject undertakes repeated non-intense attempts to exercise physical motion; the actions are often uncontrollable and fail for reasons that can be subject-internal or -external. 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 typically either physical ($t=0.176$) or metaphorical extensions of physical actions ($t=0.999$), affecting a theme/patient ($t=0.352$ and $t=0.175$, respectively). Focus is on the vainness ($t=0.962$ for vainness combined with intensity) of the durative effort ($t=0.750$ for duration adverbs).

With verbs from the cluster that contains *tščit'sja*, *pyžit'sja* and *tužit'sja*, an inanimate subject (concrete or abstract) attempts very intensely but in vain to perform what typically is a metaphorical extension of a physical action. 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.

In a nutshell, corpus-based analyses like the above single out properties that are important within a particular dataset and are likely to generalize beyond a particular dataset. However, a radial network for near synonyms expressing 'try' constructed on the basis of a linguistic data analysis alone cannot lay claim to be a truthful depiction of the mental representation of this category (cf. Sandra & Rice, 1995). Put differently, while the usage-based view of language prominent within Cognitive Linguistics places emphasis on different types of frequency effects, this does not *per se* guarantee that any of these properties are relevant to speakers of a language. The main contribution of this paper lies therefore in the attempt to validate the corpus-linguistic findings on the basis of results from experimental studies.

Exploratory Analysis

There are indications that the clustering obtained for nine near-synonymous verbs that express 'try' in Russian (see Figure 1) has a mental correlate: the results of a preliminary sorting task (Solovyev, 2006) revealed that each of the nine verbs is most often grouped together with one of the verbs it is clustered together with in the corpus-based analysis. In the following, we present additional experiments and more refined evaluation techniques to support and validate these findings.

A first exploratory sorting task (Solovyev, 2006)

Solovyev (2006; now published as Solovyev & Bajraševa, 2007) reports on a "psycho-semantic" follow-up study of Divjak and Gries (2006). Thirty-six 2nd year students of computer science at Kazan' State University in Russia received a list with the nine 'try' verbs in alphabetical order. The students were asked to sort the verbs into groups containing "words that were close in meaning". For each pair of verbs, it was calculated how often subjects had grouped them together.

Solovyev's (2006) evaluation of the results was based on a visual inspection of the co-classification matrix (cf. Table 2). He found that many students remarked they did not know the verb *tščit'sja* and had left it out of their classification. The remaining verbs clustered as follows: *norovit'* and *poryvat'sja* clustered together, as did *probovat' pytat'sja* and *pyžit'sja silit'sja* and *tužit'sja* formed a third cluster. According to Solovyev, *starat'sja* does not show any clear preference; instead, it displayed affinities with all other verbs.

In order to compare Solovyev's (2006) experimental results with our corpus-based results, and in order to homogenize the methods of evaluation across different types of experiments (see below), we designed an evaluative approach based on a point-scoring system that consists of two steps. First, we quantify the fit of the experimental results and the corpus results by means of a score. Second, we compute a random baseline to assess how likely the obtained score could have been obtained on the basis of chance alone. In what follows, we explain our evaluation method in more detail.[3]

An evaluation metric: similarity points and their baseline(s)

As mentioned above, the corpus-based analysis of the nine Russian verbs resulted in three different clusters:

- cluster 1: *poryvat'sja, norovit'* and *silit'sja*;
- cluster 2: *probovat', pytat'sja,* and *starat'sja*;
- cluster 3: *pyžit'sja, tščit'sja* and *tužit'sja.*

In order to quantify the convergence between the corpus-based cluster solution and the results of the sorting task, we generated a co-classification matrix. Each cell of this matrix provides the frequency with which the verb listed in the row has been sorted together with the verb in the corresponding column. Table 2 provides this matrix for the data discussed in Solovyev (2006).[4]

This symmetric matrix has an unpopulated main diagonal since each verb *v* is by definition sorted into a group with itself. Second, in order to avoid basing our conclusions on raw frequencies of occurrence only, we computed for each cell (i) its expected frequency (according to the formula in (1)) and (ii) its Pearson residual (according to the formula in (2));[5] positive versus negative Pearson residuals indicate that a particular frequency is higher or lower than expected by chance.

(1) expected frequency $= \dfrac{total_{row} \cdot total_{column}}{total_{table}}$

(2) Pearson residual $= \dfrac{observed - expected}{\sqrt{expected}}$

Computing all Pearson residuals for the data presented in Table 2 results in Table 3; the bold-typed figures in Table 3 highlight the row-wise maxima.

Table 2. Co-classification matrix (data from Solovyev, 2006)

	norovit’	*poryvat’sja*	*silit’sja*	*probovat’*	*pytat’sja*	*starat’sja*	*pyžit’sja*	*tščit’sja*	*tužit’sja*
norovit’		17	3	7	4	8	1	2	3
poryvat’sja	17		2	9	6	3	2	0	1
silit’sja	3	2		2	8	10	20	5	21
probovat’	7	9	2		23	5	0	0	1
pytat’sja	4	6	8	23		10	2	1	2
starat’sja	8	3	10	5	10		4	1	7
pyžit’sja	1	2	20	0	2	4		7	27
tščit’sja	2	0	5	0	1	1	7		5
tužit’sja	3	1	21	1	2	7	25	5	

Table 3. Pearson residuals for the co-classification matrix in Table 1

	norovit’	*poryvat’sja*	*silit’sja*	*probovat’*	*pytat’sja*	*starat’sja*	*pyžit’sja*	*tščit’sja*	*tužit’sja*
norovit’		**6.55**	-1.52	1.08	-0.66	1.49	-2.05	-0.06	-1.36
poryvat’sja	**6.55**		-1.7	2.39	0.48	-0.6	-1.46	-1.36	-1.98
silit’sja	-1.52	-1.7		-1.97	-0.26	0.91	3.39	0.95	**3.4**
probovat’	1.08	2.39	-1.97		**7.14**	0.01	-2.51	-1.47	-2.21
pytat’sja	-0.66	0.48	-0.26	**7.14**		1.68	-2.01	-0.99	-2.13
starat’sja	1.49	-0.6	0.91	0.01	**1.68**		-0.96	-0.82	0.05
pyžit’sja	-2.05	-1.46	3.39	-2.51	-2.01	-0.96		2.49	**5.5**
tščit’sja	-0.06	-1.36	0.95	-1.47	-0.99	-0.82	**2.49**		1.15
tužit’sja	-1.36	-1.98	3.4	-2.21	-2.13	0.05	**5.5**	1.15	

Next, we computed a point score that quantifies how well the sorting data fit the corpus data. To give an example, the high Pearson residual in the first row of Table 3 reflects that *norovit'* was sorted together with *poryvat'sja* much more often than expected by chance; this is reflected in the following scoring system:

- if a target verb's highest Pearson residual in the sorting data was observed for a verb that was assigned to the same cluster as the target verb belongs to in the corpus data, this scored one point;
- if a target verb's highest Pearson residual in the sorting data was observed for a verb that was assigned to another cluster than the target verb belongs to in the corpus data, this scored zero points.

From Table 3, it is clear that the minimum and the maximum scores that can be obtained with our scoring system are 0 and 9 points respectively. Since all verbs except for *silit'sja* have their highest Pearson residual for another verb from the same corpus-based cluster, we scored 8 points, which is a very good result. To test whether our result is also sufficiently — i.e., significantly — different from chance, we adopted a simulation-based approach. We first enumerated all scores any verb could theoretically obtain. Since each verb is part of a three-verb cluster, it could theoretically score 1 for either of the two verbs from the same cluster or 0 for any of the six remaining verbs. Thus, each verb will on average contribute a score of $^2/_8$ to the overall point score, and the overall expected score will be 2.25. This result indicates that our score of 8 is 3.5 times higher than expected by chance. To test this outcome for significance while avoiding a computationally intensive permutational test, we used a bootstrapping approach. We generated a vector with all possible scores {1, 1, 0, 0, 0, 0, 0, 0}, sampled one value from this vector nine times (once for each verb) with replacement, and added these nine values up to one sample sum. We did this 100,000 times, obtaining 100,000 sums. The frequency distribution of these sums is summarized in Table 4, showing the most important quantiles resulting from the simulation. Since the observed sum of 8 is not even attested in the frequency range covering 99.5% of the data, the score of 8 observed in the real, non-simulated data is significantly higher than expected by chance. More precisely, the number of times the sample sum was 8 (our observed value) or higher was 12 out of all 100,000 times: $p_{\text{one-tailed}} = 0.00012$.

In sum, the results of Solovyev's (2006) sorting experiment support the three cluster-solution suggested by the corpus analysis. Admittedly, Solovyev, elicited sortings in a rather crude way, i.e. without providing the intended syntactic and

Table 4. Quantiles from the Simulation

Quantile	0.005	0.01	0.025	0.05	0.5	0.95	0.975	0.99	0.995
Value	0	0	0	0	2	4	5	6	6

semantic context for the verbs. In section *Three Sorting Tasks*, we discuss the results from our own sorting experiments; section *A Gap-filling Task* discusses a second experimental validation in the form of a gap-filling task.

Two Experiments[6]

In this section, we aim to provide an answer to two interrelated questions concerned with the degree to which the corpus-based results are corroborated by experimental evidence and the degree to which corpus-based studies contribute to linguistic investigations of semantic and conceptual issues. First, do native speakers in contextually more controlled experiments produce clusters that resemble the clustering obtained from the analysis of contextually rich corpus data? If they do, this would illustrate the strength of a corpus-based approach in general, and the importance of distributional aspects and similarity for the acquisition and mental organization of lexemes in particular. If they do not, this could suggest that native speakers think of synonymy as a lexical relation holding between words in pairs only or, of course, that the corpus data have little to contribute to how speakers distinguish synonymous words. Second, are native speakers sensitive to the properties that, on the basis of corpus data, are claimed to be strongly associated with a cluster of verbs (cf. Arppe & Järvikivi, 2007; for the Finnish synonym pair *miettiä* and *pohtia* meaning 'think'), or is it only with reference to discourse-pragmatic and socio-variationist properties that reliable distinctions between near synonyms can be made?

Before embarking on the analysis, one caveat is in order. Whenever reference is made to "cognitive reality", no position is taken as to the exact mental representation or processing of lexical clusters. In our view, our results suggest that information about distinctive properties as they fall out from a corpus-driven linguistic analysis is likely to be stored, and whichever way that lexical information is stored, it is very well suited to produce clusters that are correlated with, or fall out from, distributional characteristics (cf. section *Conclusion* for more discussion).

Three Sorting Tasks

Experimental design. 46 third-year IT students from the Department of Computer Science and Economics at the Moscow Steel and Alloys Institute (http: //www.misis.ru) in Russia, were presented with a questionnaire that contained instructions for three sorting tasks.[7] In each task, participants were presented with nine sentences that differed only with respect to the main verb expressing 'try' that was used. The carrier sentence and its translation is given in (3); the underlined gap was filled by past tense forms of the nine different verbs meaning 'try' in Russian. There was no indication in the corpus data that the chosen subject *cripple* and the selected activity *walk*, would favor a particular verb or group of verbs.

(3) a. После операции калека ______________ ходить без помощи костылей.

b. After the operation, the cripple <u>tried </u>to walk without the help of crutches.

In task 1, the participants were asked to sort the nine sentences into any number of groups such that sentences they thought were more similar to each other ended up in the same group, while sentences that were found to be less similar to each other were sorted into different groups. The subjects were asked to indicate the grouping by assigning identical numbers, letters or symbols to sentences they thought belonged to the same group.

In task 2, the subjects were asked to revisit the same sentences and sort them into three groups such that sentences they thought were more similar to each other were sorted into the same group while sentences that were less similar to each other were sorted into different groups; again, the subjects indicated their groupings with numbers, letters or symbols.

In task 3, the subjects were asked to revisit the same sentences, but this time to sort them into three groups containing three verbs each on the basis of the same criteria.

In other words, the three tasks systematically narrowed down the options for possible sortings, offering us different standards of comparison for our corpus-based results, as will be discussed in the following section.

Results from hypothesis-testing. The data were evaluated in the same way as Solovyev's (2006) data. For each verb in each task, we counted how often it was sorted into the same group as each other verb and computed the Pearson residuals of the resulting co-classification matrix. The resulting matrices are provided in Tables 5–7 for tasks 1–3, respectively.

The point score resulting from each of these tables is 8: in the experiments, all verbs but *silit'sja* (for an explanation see below) prefer to be grouped with verbs from the same cluster they were associated with in our previous corpus-based clustering solution.

For each of these three sets of results, we computed the same simulation as presented above for Solovyev's (2006) data. In all three cases, the results were identical. For all tasks, a point score of 8 or higher was obtained only 12 times out of

Table 5. Pearson residuals for the co-classification matrix of Task 1

	norovit’	*poryvat’sja*	*silit’sja*	*probovat’*	*pytat’sja*	*starat’sja*	*pyžit’sja*	*tščit’sja*	*tužit’sja*
norovit’		**5.7**	-2.27	-1.5	-2.12	-2.18	-2.56	-0.75	-2.63
poryvat’sja	**5.7**		-3.22	-1.45	-1	-0.54	-3.04	-1.59	-3.36
silit’sja	-2.27	-3.22		-1.67	-2.25	-1.84	1.73	0.15	**2.74**
probovat’	-1.5	-1.45	-1.67		**3.77**	1.32	-2.93	-2.9	-3
pytat’sja	-2.12	-1	-2.25	**3.77**		3.22	-3.26	-2.97	-3.32
starat’sja	-2.18	-0.54	-1.84	1.32	**3.22**		-2.32	-2.73	-2.64
pyžit’sja	-2.56	-3.04	1.73	-2.93	-3.26	-2.32		0.19	**4.39**
tščit’sja	-0.75	-1.59	-0.15	-2.9	-2.97	-2.73	0.19		**0.36**
tužit’sja	-2.63	-3.36	2.74	-3	-3.32	-2.64	**4.39**	0.36	

Table 6. Pearson residuals for the co-classification matrix of Task 2

	norovit’	*poryvat’sja*	*silit’sja*	*probovat’*	*pytat’sja*	*starat’sja*	*pyžit’sja*	*tščit’sja*	*tužit’sja*
norovit’		**4.22**	-2.36	-0.09	-0.39	-0.74	-2.25	-2.54	-2.76
poryvat’sja	**4.22**		-1.96	-0.86	-0.65	-0.53	-2.83	-1.87	-3.11
silit’sja	-2.36	-1.96		-1.51	-1.55	-0.98	0.58	0.07	**1.45**
probovat’	-0.09	-0.86	-1.51		**2.7**	2.18	-3.04	-3.38	-3.07
pytat’sja	-0.39	-0.65	-1.55	**2.7**		2.23	-3.24	-2.8	-2.36
starat’sja	-0.74	-0.53	-0.98	2.18	**2.23**		-2.92	-2.97	-2.73
pyžit’sja	-2.25	-2.83	0.58	-3.04	-3.24	-2.92		3.05	**4.22**
tščit’sja	-2.54	-1.87	0.07	-3.38	-2.8	-2.97	**3.05**		1.96
tužit’sja	-2.76	-3.11	1.45	-3.07	-2.36	-2.73	**4.22**	1.96	

Table 7. Pearson residuals for the co-classification matrix of Task 3

	norovit’	*poryvat’sja*	*silit’sja*	*probovat’*	*pytat’sja*	*starat’sja*	*pyžit’sja*	*tščit’sja*	*tužit’sja*
norovit’		**4.2**	1.42	-1.69	-2.18	-1.49	-2.2	-0.12	-2.68
poryvat’sja	**4.2**		-1.07	-2.11	-3.12	-1.39	-2.88	1.28	-2.86
silit’sja	1.42	-1.07		-1.88	-2.11	-2.43	**1.69**	-2.09	1.47
probovat’	-1.69	-211	-1.88		**4.75**	3.61	-3.68	-3.14	-3.67
pytat’sja	-2.18	-3.12	-2.11	**4.75**		3.9	-3.16	-2.61	-3.39
starat’sja	-1.49	-1.39	-2.43	3.61	**3.9**		-2.95	-3.43	-3.44
pyžit’sja	-2.2	-2.88	1.69	-3.68	-3.16	-2.95		0.45	**5.01**
tščit’sja	-0.12	1.28	-2.09	-3.14	-2.61	-3.43	0.45		**1.76**
tužit’sja	-2.68	-2.86	1.47	-3.67	-3.39	-3.44	**5.01**	1.76	

Table 8. Quantiles from the Three Simulation Tasks

Quantile	**0.005**	**0.01**	**0.025**	**0.05**	**0.5**	**0.95**	**0.975**	**0.99**	**0.995**
Task 1–3	0	0	0	0	2	4	5	6	6

all 100,000 simulation runs; thus $p_{\text{one-tailed}} = 0.00012$. See Table 8 for the quantiles associated with each task's simulation.

Thus, we find that the subjects — regardless of the exact sorting instructions they were given — strongly prefer sorting solutions that are consistent with the corpus-based clustering. Throughout, the point scores obtained are 3.5 times higher than expected by chance, and that ratio difference is highly significant according to all Monte Carlo simulations. Overall, eight out of nine verbs are grouped with verbs from the cluster they were assigned to in the corpus-based analysis. Across tasks, seven of the nine verbs are classified identically: *tščit'sja* changes between *pyžit'sja* in sorting task one and *tužit'sja* in tasks two and three, but stays within its corpus-based cluster, whereas *silit'sja* transgresses cluster boundaries in all three tasks, clustering with *pyžit'sja* in task three and with *tužit'sja* in tasks one and two. A possible cause for this divergence is the absence of pragmatic variables in the behavioral profile: just like *pyžit'sja* and *tužit'sja*, *silit'sja* strongly foreshadows failure of the attempted action.[8]

Results from further cluster-analytic exploration. Additional confirmation for the existence of three clusters in the elicited data, clusters that strongly resemble those found in the corpus data, stems from computing cluster analyses on each of the co-classification matrices from tasks 1 through 3. We computed three hierarchical agglomerative cluster analyses — one on the co-classification matrix of each task — and, in order to rule out methodological artifacts, we used the same settings that Divjak and Gries (2006) used for their corpus data (similarity measure: Canberra, amalgamation rule: Ward). In what follows, we will briefly discuss the results from each clustering to see how the results relate to our earlier corpus-based results.

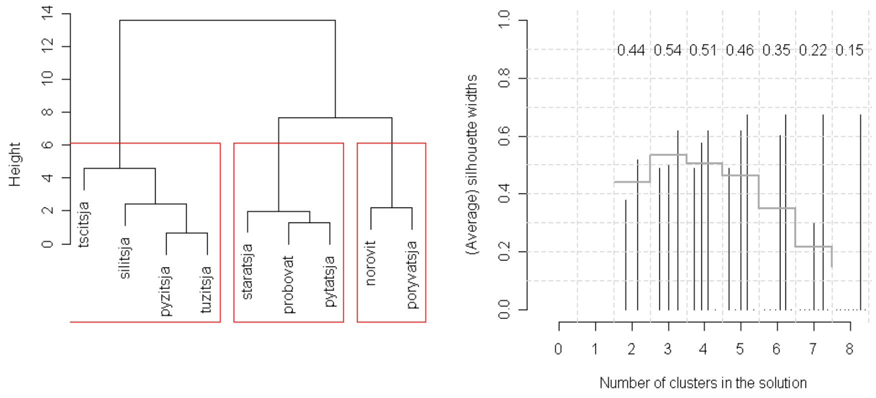

Figure 2. Cluster analysis for Task 1 of our sorting experiment.

The cluster analysis on the data from the first sorting task yielded the results represented in Figure 2.

For this cluster analysis, we adopted a three-cluster solution (as shown in the left panel) for three reasons. First, the average of all silhouette widths reaches its maximum when three clusters are assumed (as shown in the right panel). Second, a *k*-means cluster analysis and a linear discriminant analysis on the basis of the three-cluster solution could reproduce the clustering perfectly. Third, with one exception, all *F*-values computed for each cluster are smaller than 1, thus supporting the assumption that a three-cluster solution results in homogeneous groups.

The cluster analyses on the data from the second and third sorting tasks yielded the results represented in Figure 3 and Figure 4.

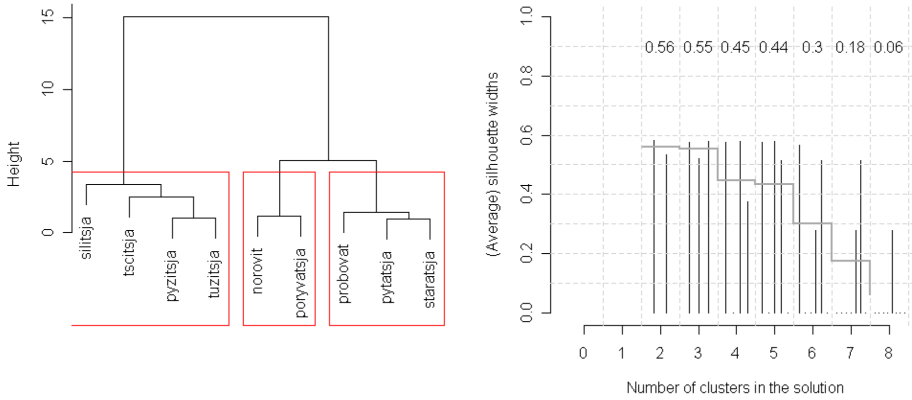

Figure 3. Cluster analysis for Task 2 of our sorting experiment.

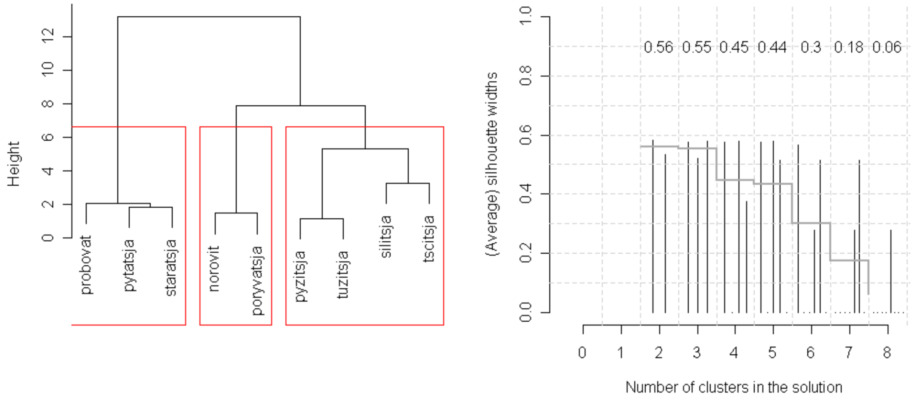

Figure 4. Cluster analysis for Task 3 of our sorting experiment.

In both cases, a three-cluster solution (as shown in the left panel) and a two-cluster solution are about equally likely. While the average of all silhouette widths reaches its maximum when two clusters are assumed (as shown in the right panel), the difference to the average silhouette width for a three-cluster solution is negligible. Also, *k*-means cluster analyses and linear discriminant analyses for both the two and three-cluster solutions reproduced the clustering perfectly, and the *F*-values for both clustering solutions reflected the same degree of homogeneity. Given the equality of the results and the significant scoring point results, the data can, therefore, be considered compatible with the corpus-based solution.

In sum, in each task all verbs but *silit'sja* ended up in the same cluster as in the corpus data (again, for an explanation see below). We take this result as strong evidence for the compatibility of the experimental and the corpus-based clusterings. More rigorously, we computed Fowlkes and Mallows's (1983) measure of association for comparing two hierarchical cluster solutions, B_k in order to quantify the degree of fit between the dendrograms based on the corpus data on the one hand and the three dendrograms based on the three sorting tasks on the other hand; in each case, we obtained a high value of 0.74.

A Gap-filling Task

Experimental design. In addition to the above sorting experiment, we performed a gap-filling experiment (similar to the one employed by Dąbrowska, in press) to check whether there was a quantitative dimension to the ID tag levels that had been singled out in the corpus-based study as highly distinctive for clusters using *t*-scores. Arguably, the *t*-values resulting from cluster analysis give a rough corpus equivalent of the probabilistic notion of cue validity from the domain of categorization studies. A feature *f* has high cue validity for category *c* if most members of *c* exhibit *f* and most non-members of *c* lack *f*. Similarly, a high *t*-value for a feature *f* linked with a cluster signals a strong association of that particular feature with that particular cluster, and less so with other clusters. In other words, in both cases high values signal highly distinctive properties.

Subjects were presented with a questionnaire containing a list of 27 verbs (each of the nine 'try' verbs three times) as well as 27 sentences. The 27 sentences were taken directly from the Russian dataset on which the corpus analysis was based and were chosen such that each carrier sentence exhibited particularly high *t*-values for the 'try' verb that would then be deleted and replaced with a gap. Note that, since the ID tags only capture sentence-internal properties, the experimental carrier sentences are not impoverished clones of, or stripped down substitutes for, real-life corpus sentences — rather, if the BP approach is on the right track, the ID tags with high *t*-values should be strong and context-independent cues to the verbs. A detailed enumeration of these properties was provided in section *Evaluating the results*, so we will limit ourselves here to summarizing the main ID tags per cluster.

The cluster that contains *probovat'*, *pytat'sja* and *starat'sja* is defined by the combined strongest ID tags as applying to human beings that are exhorted to undertake an attempt to carry out a physical action, to move others or to undertake motion in the figurative sense; often, these activities are negated. The three 'try' verbs are typically used as main verb in a main clause. The cluster with *silit'sja*, *norovit'* and *poryvat'sja* seems reserved for situations in which an inanimate subject (concrete or abstract) attempts for a certain amount of time, very intensely but in vain, to perform what typically are physical activities or metaphorical extensions of physical actions. Finally, in trying as expressed by *tščit'sja*, *pyžit'sja* and *tužit'sja*, an inanimate subject undertakes repeated non-intense attempts to exercise physical motion; the actions are often uncontrollable and fail because of internal/external reasons. These three verbs, in particular, are often used as participles.

The questionnaires were presented to 45 third-year IT students from the Department of Computer Science and Economics at the Moscow Steel and Alloys Institute (http://www.misis.ru) in Russia, who were asked to fill the gaps with the verbs from the list (cf. (4) for an example). For their convenience, each verb was listed three times.[9]

(4) Раньше он, наверное, ________________ бежать, но теперь понял, что от этого сутулого человека никуда не убежишь.

Earlier he would, probably, _____________ to run, but now he understood that you can't run anywhere from this stooping man.

Results from hypothesis-testing. Since we employed the same kind of test for both experimental studies, the characterization of the corresponding test can now be abbreviated. In the gap-filling experiment, subjects were provided with a carrier sentence from which the verb meaning 'try' that was used in the corpus example had been deleted and were asked to enter the verb (chosen from all nine verbs) they considered most fitting. By analogy to the above procedure, we therefore began by generating a gap-filling preference matrix, each cell of which provides the frequency with which the (stimulus) verb listed in the row has resulted in the gap-filling verb from the column. Table 9 provides this gap-filling preference matrix.

This matrix is *not* symmetric, and this time its main diagonal is populated as we hypothesize that each stimulus verb should have triggered the verb that was used in the sentences originally or a verb that belongs to the same cluster. Second, we computed each cell's Pearson residual in the same way as above and provide all Pearson residuals for Table 9 in Table 10.

Table 9. Gap-filling preference matrix

	Response								
Stimulus	*norovit'*	*poryvat'sja*	*silit'sja*	*probovat'*	*pytat'sja*	*starat'sja*	*pyžit'sja*	*tščit'sja*	*tužit'sja*
norovit'	59	30	2	5	9	4	4	6	6
poryvat'sja	16	42	19	10	11	4	3	18	9
silit'sja	8	13	28	6	8	2	16	18	31
probovat'	9	14	11	35	28	10	7	3	14
pytat'sja	4	6	17	24	8	7	26	18	16
starat'sja	0	1	4	34	15	53	5	5	8
pyžit'sja	7	4	8	5	20	22	35	20	13
tščit'sja	19	21	22	6	18	20	3	13	11
tužit'sja	12	5	20	3	20	14	17	27	13

The third step, again, consists of computing a point score that quantifies how well the corpus data fit the gap-filling preferences. As before, a high Pearson residual (in Table 10) reflects that one verb was much more often provided as a gap-filler for another verb. Yet, in this experiment, there is a third scoring option, namely the deleted stimulus verb being the same as the gap-filling verb provided by the subject. We therefore adopted the following scoring system:

- when a stimulus verb's highest Pearson residual was observed for the same verb as a gap-filler, this scored two points;
- when a stimulus verb's highest Pearson residual was observed for a verb that was in the same cluster in the corpus data, this scored one point;
- when a stimulus verb's highest Pearson residual in the sorting data was not observed for a verb that was in the same cluster in the corpus data, this scored zero points.

Table 10. Pearson residuals for the gap-filling preference matrix in Table 8

	Response								
Stimulus	*norovit'*	*poryvat'sja*	*silit'sja*	*probovat'*	*pytat'sja*	*starat'sja*	*pyžit'sja*	*tščit'sja*	*tužit'sja*
norovit'	**11.78**	4.04	-3.21	-2.35	-1.48	-2.77	-2.39	-2.08	-1.93
poryvat'sja	0.22	**6.79**	1.09	-1.18	-1.14	-2.9	-2.79	0.93	-1.27
silit'sja	-1.79	-0.55	3.51	-2.19	-1.86	-3.38	0.86	0.99	**4.77**
probovat'	-1.56	-0.32	-0.97	**5.44**	3.22	-1.35	-1.67	-3	0.11
pytat'sja	-2.75	-2.27	0.76	2.74	-1.77	-2.01	**3.81**	1.12	0.81
starat'sja	-3.79	-3.55	-2.68	5.48	0.09	**10.07**	-2.11	-2.35	-1.38
pyžit'sja	-2.14	-2.94	-1.82	-2.53	1.08	1.62	**5.94**	1.38	-0.24
tščit'sja	0.95	1.4	**1.83**	-2.25	0.6	1.14	-2.81	-0.42	-0.75
tužit'sja	-0.78	-2.63	1.38	-3	1.18	-0.32	1.1	**3.33**	-0.16

As before, the bold-typed figures in Table 10 correspond to the row-wise maxima. It is clear from the table that we score 11 points out of the range of possible scores from 0 to 18. To test whether this result is significantly different from chance, we first note down all possible scores any verb could obtain. Since each verb is part of a three-verb cluster, this means that each verb could theoretically score

- 2 points if it most strongly preferred itself as a gap-filler;
- 1 point for either of two verbs from the same cluster;
- 0 point for any of the six remaining verbs.

Thus, each verb will on average contribute a score of $^4/_9$ to the overall point score and the overall expected score will be 4. To test the difference between our obtained 11 and the expected 4 points for significance, we generated a vector with all possible scores {2, 1, 1, 0, 0, 0, 0, 0, 0}, sampled (with replacement) nine values from this vector (one for each verb), and added these nine values up to one sample sum. We did this 100,000 times and then computed the number of times the sample sum was 11 (our observed value) or higher. This was the case in 251 out of 100,000 times; thus, $p_{\text{one-tailed}} = 0.00251$, which shows that the observed value of 11 is not only 2.75 times higher than expected by chance, but also very significantly so. In addition, we provide some quantiles resulting from the simulation in Table 11.

Table 11. Quantiles from the Simulation

Quantile	**0.005**	**0.01**	**0.025**	**0.05**	**0.5**	**0.95**	**0.975**	**0.99**	**0.995**
Value	0	0	0	1	4	8	8	9	10

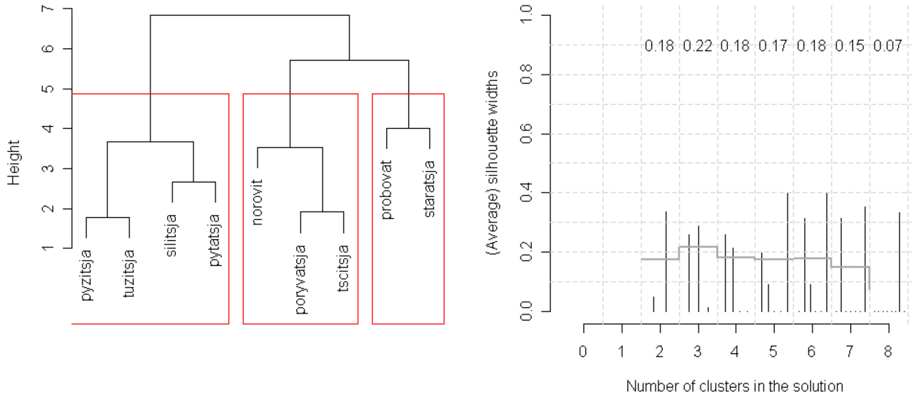

Figure 5. Cluster analysis for our gap-filling experiment.

In sum, the results from our gap-filling experiment correlate well with the results of the clusters that were arrived at on the basis of the corpus data. This in turn supports the BP approach speakers are very sensitive to the ID tags and contextual clues that were provided in the experiment and that are at the heart of the BP approach.

Results from further cluster-analytic exploration. As before in section *Results from further cluster-analytic exploration*, we also computed a hierarchical cluster analysis on the data from the gap-filling experiment. Consider Figure 5 for the dendrogram.

For this cluster analysis, we adopted a three-cluster solution (as shown in the left panel) for three reasons. First, the average of all silhouette widths reaches its maximum when three clusters are assumed (as shown in the right panel). Second, a *k*-means cluster analysis and a linear discriminant analysis on the basis of the three-cluster solution could reproduce the clustering nearly perfectly (88.89% classification accuracy in the *k*-means clustering, 100% classification accuracy in the LDA). Third, all but two *F*-values computed for each cluster are smaller than 1, thus supporting the assumption that a three-cluster solution results in homogeneous groups. However, the results obtained by comparing the cluster trees from the gap-filling experiment to the corpus data are not quite as supportive as those from the sorting experiment: Fowlkes and Mallows's (1983) measure of association B_k for the fit between the clustering of the gap-filling task and the corpus-based clustering of section *Tagging for Meaning* equals only 0.32. This should not come as a surprise, however. The sorting data stem from an experimental design that is free of noise and uncontrolled variation as each stimulus sentence only differed with respect to the main 'try' verb under consideration. In the gap-filling task, however, each stimulus sentence was selected to represent a particular set of *t*-values that had proven to be relevant in the corpus-based clustering solution. Since we wanted to chose authentic sentences, each sentence also contains a variety of additional ID tags, which results in additional associations to (verbs from) other clusters. Thus, while the *t*-values according to which we selected the stimuli do result in the hypothesized gap-filling patterns on the whole, the results for the gap-filling experiment are not as clean as those for the sorting data.

Conclusions

Synonym clusters "exist" in corpora and mind. Our findings reveal that the corpus-based representation of the synonym clusters (proposed in Divjak and Gries, 2006) is not a by-product of corpus composition or of the statistical technique used (in the sense that a cluster analysis will always output some sort of structure).

Instead, there seems to be a psychological reality corresponding to clusters of near synonyms. We provide evidence that, in terms of its predictive power, a corpus-based cluster structure is a fair approximation of the mental representation of the categories in question. Our study thus has relevant descriptive, methodological, and theoretical implications for the field of cognitive semantics.

First of all, the present findings confirm that the verbs expressing 'try' in Russian can be divided into three fairly well distinguishable clusters. As such, the sorting results provide additional support for the semantic analysis of the nine verbs outlined in Divjak and Gries (2006). This conclusion is reinforced by the fact that the gap-filling experiment revealed the discriminatory power of the ID tag levels with high *t*-values on which Divjak and Gries (2006) based their analysis. Although the strong correspondence of the experimental results and the corpus data might fit some other semantic interpretation of the main meaning of the clusters, the present results are, at the very least, highly compatible with the semantic account presented. On a more abstract level, the results show that speakers group these near synonyms into clusters, not pairs. Hence, near synonymy is not about pairs of words that entertain dichotomous, dyadic relations (as assumed in the structuralist era — see Quine (1964) for an early reaction against this view), but about groups of words that are more similar to each other than to (words belonging to) other groups of semantically similar words. Although to the best of our knowledge no theoretical importance has been attached to analyzing synonyms as pairs, it is surprising that (cognitive) linguists would consistently (albeit implicitly) represent a lexical phenomenon, synonymy, in a way that lacks any cognitive underpinning. With the notable exception of Edmonds and Hirst (2002), many if not most analyses we are aware of tend to analyze synonyms in a pairwise fashion; compare here standard textbook references (e.g., Cruse, 1986; Saaed, 1997), lexical-semantic studies (e.g., Geeraerts, 1985; Mondry and Taylor, 1992) as well as corpus-based studies (cf., e.g., Gries, 2003; Kjellmer, 2003; Taylor, 2003), etc.

From a methodological perspective, too, our findings are of importance: the results of both experiments correspond (significantly) to the results of the corpus-based BP approach. Subjects' behavior strongly suggests that they have at least some knowledge of the overall similarities between the nine near synonyms, a similarity that appears to be adequately captured by the BPs for the nine verbs: not only did the subjects sort the nine near synonyms into groups that correspond to the corpus-derived clusters, intersubstitutability between verbs from different semantic corpus-based clusters also proved to be rather low. Subjects are likewise sensitive to a corpus-based operationalization of cue validity as they fill gaps as predicted by the distributional features of the stimulus sentences. Thus, a corpus-based approach to language description, and the BP approach in particular, receives strong experimental support: significant (yet not necessarily sufficient) components of "meaning", and maybe even of the way in which verbs are stored and/or processed, can be extracted by studying usage in (textual) context. If used properly, corpus data provide reliable access to linguistic knowledge, as is proven by the high "cue-validity" of (generalizations over) properties selected on the basis of corpus research. Likewise, psycholinguists might want to worry less about length-based, familiarity-based, or frequency-based lexical effects and more about factors such as the ID tags and levels we identified in Table 1 (which take into account some several dozen properties ranging from inflectional to collocational behavior) that might affect experimental results when working with sentential stimuli.[10]

The question remains as to how the match arises between the corpus-based distributional findings and the experimentally-observed preferences. In our view, our results provide additional support for an exemplar-based conception of the acquisition and representation of language that is alluded to in the epigraph of this paper and for which the supporting body of evidence is growing. In a way similar to Hoey (2005),[11] Dąbrowska (in press), for example, proposes that learners acquire the meanings of words on the basis of contextual and distributional cues provided in usage events by (i) storing lexically-specific knowledge of semantic and collocational preferences and (ii) forming more phonologically and semantically abstract generalizations or schemas on the basis of recurrent exposure to particular components of meaning. In other words, in line with recent work on exemplar theory (cf. Bybee, 2000; Pierrehumbert, 2001) and Abbot-Smith and Tomasello (2006, p. 275) we are inclined to argue for a 'hybrid' view

> […] in which acquisition depends on exemplar learning and retention, out of which permanent abstract schemas gradually emerge and are immanent across the summed similarity of exemplar collections. These schemas are graded in strength depending on the number of exemplars and the degree to which semantic similarity is reinforced by phonological, lexical, and distributional similarity.

Applied to our verbs, this hybrid view implies that the acquisition of verbs expressing 'try' involves memorizing instances — in multidimensional syntactic-semantic space represented as a dot — as a "cloud" of exemplars. Whenever a speaker encounters yet another instance of one of the nine verbs meaning 'try', the memory representation of these verbs and their actual uses is updated with the information contained in the most recent usage event. However, not all actual instances need be remembered: memory traces may decay over time and while particular salient usage events may remain accessible, what remains for the most part may well be generalizations based on many similar but now forgotten usage events. These generalizations are assumed to involve probabilistic knowledge of distributional patterns (in this case the combination of semantic properties of agent, activity, adverb, but also grammatical co-occurrences or colligations) that in our approach correspond to the ID tag levels characterized by high *t*-values for verbs in semantically fairly homogeneous clusters.

On this view, the results of the sorting and the gap-filling task would result from subjects accessing traces of memory representations for the use of the verbs. More specifically, the contextual clues provided in the gap-filling task facilitate access of a particular sub-region of the syntactic-semantic space containing a cloud of traces for verbs that were used in a similar way. The likelihood that subjects produce the same or a similar verb thus increases strongly. The strong similarity between the corpus-based and the experimental results is due to the BP approach tapping into exactly those distributional patterns that help shape the arrangement of verbs in syntactic-semantic space.

In sum, the corpus-based BP approach is an objective, data-driven alternative to intuitive approaches to semantics with at least two major advantages. On the one hand, it yields descriptions at a previously not utilized level of precision and makes it possible to answer notoriously difficult questions in the domains of polysemy, near synonymy, and lexical fields (cf. Gries, 2006; Dąbrowska, in press; Divjak, 2006; Divjak & Gries, 2006) including issues like network construction, prototype identification, and the analysis of similarities of words and word senses (i.e., the structure of word senses and lexical fields). On the other hand, it correlates strongly with different experimental methods: sorting and gap-filling (cf. above and Dąbrowska, in press), sentence elicitation and video descriptions (cf. again Dąbrowska, in press), and forced-choice selection and judgment tasks (cf. Arppe & Järvikivi, 2007). We therefore hope that, as more and more diverse corpora become available, a combined method of investigation will be more frequently applied within cognitive (lexical) semantics. Corpora clearly have the potential to yield excellent hypotheses that can be subjected in a straightforward way to experimental verification and, in the case of evidence as converging as in the present study, strengthen our account of linguistic phenomena as elusive as lexical semantics.

Notes

* Many thanks to Ewa Dąbrowska, Antti Arppe and Stefanie Wulff for commenting on an earlier draft of this article. Dagmar Divjak's contribution to the research reported on in this paper was funded jointly by the University of Sheffield and a F.W.O. Vlaanderen (honorary) postdoctoral scholarship.

[1] The *t*-values of an ID tag percentage *p* for a cluster *c* out of *n* clusters are computed as follows: (mean (*p* within *c*) — mean (*p* across all *n* clusters)) ÷ standard deviation *p* across all *n* clusters. However, given the small number of elements, we use the *t*-values only descriptively and not for

the purpose of performing significance tests (cf. Backhaus, Erichson, Plinke, & Weiber, 2003, p. 310–2).

2. The absolute values of these *t*'s may well seem very low, but this is expected given that we are dealing with near-synonymous verbs, verbs that are, by definition, highly similar in meaning. If the *t*-values had been large, we would have had reason to doubt that these verbs actually belonged to the same semantic group, let alone to the same cluster of synonymous verbs.

3. All computations were done with R for Windows 2.6.1 (R Development Core Team, 2007).

4. We thank Valerij Solovyev for making his data available to us.

5. We did not use standardized Pearson residuals because the residuals are only used for within-row comparisons.

6. The experiments were approved by the University of Sheffield institutional review board. We thank Leonid Oknyansky for assisting us in constructing the experimental materials and Katya Chown for double-checking the instructions given in the questionnaires.

7. We thank Andrej Kibrik and Vladimir Polyakov for their help in carrying out the experiments.

8. The absence of socio-lectal factors can hardly have played any role as *pyžit'sja* and *tužit'sja* are consistently labelled "colloquial" or even "vulgar" in dictionaries, whereas *silit'sja* is not.

9. For the cluster [*probovat'/pytat'sja/starats'ja*], example sentences were selected that contained an animate subject and a physical action, a motion activity that contained an "other" or figurative motion. For [*silit'sja/poryvat'sja/norovit'*], subjects were inanimate and carried out physical motion. For [*tscit'sja/pyžit'sja/tužit'sja*], an inanimate subject/group/institution undertook a physical activity that included an "other", literally or figuratively.

10. We thank one of the anonymous reviewers for this remark.

11. Dąbrowska (in press) investigates how the meanings of rare verbs of walking or running such as, e.g., *stagger*, *hobble*, *plod*, or *saunter,* are acquired. In two case studies, she shows that verbs are, firstly, reliably associated with semantic and collocational preferences of the main arguments and complements of the verbs and, secondly, that speakers use contextual and referential knowledge to identify which of a set of semantically similar verbs is most appropriate in a given context or a for a particular scenario.

References

Abboth-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review, 23*, 275–90.

Apresjan, Ju. D. (1995). *Избранные труды. Том I. Лексическая семантика: синонимические средства языка.* [*Selected works. Volume I. Lexical semantics: The synonymic means of laguage*]. Moskva: Škola "Jazyki Russkoj Kul'tury"

Arppe, A., & Järvikivi, J. (2007). Every method counts — Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory, 3*(2), 131–59.

Atkins B. T. (1987) "Semantic ID Tags: corpus evidence for dictionary senses", In Proceedings of the Third Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary The Uses of Large Text Databases, Waterloo, Canada, pp. 17–36.

Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2003). *Multivariate Analysemethoden: eine anwendungsorientierte Einführung*. [*Methods of multivariate analysis an application-focussed introduction*]. 10th ed. Berlin, Heidelberg, New York: Springer.

Bybee, J. (2000). The phonology of the lexicon: evidence from lexical diffusion. In M. Barlow & S. Kemmer (Eds.), *Usage-based Models of Language* (pp. 65–85). Stanford: CSLI Publications.

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

Dąbrowska, E. (in press). Words as constructions. In V. Evans & S. Pourcel (Eds.), *New Directions in Cognitive Linguistics.* Amsterdam and Philadelphia: John Benjamins.

Divjak, D. (2006). Ways of intending: A corpus-based cognitive linguistic approach to near synonyms in Russian. In St. Th. Gries & A. Stefanowitsch. (Eds.), *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis* (pp. 19–56). Berlin, Heidelberg, New York: 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.

Edmonds, Ph., & Hirst, G. (2002). Near synonymy and lexical choice. *Computational Linguistics, 28*(2), 105–44.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association, 78*(383), 553–69.

Geeraerts, D. (1985). Preponderantieverschillen bij bijna-synoniemen. [*Preponderance differences in near-synonyms*]. *De Nieuwe Taalgids, 78*, 18–27.

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

Gries, St.Th. (2003). Testing the sub-test: A collocational overlap analysis of *-ic* and *-ical* adjectives. *International Journal of Corpus Linguistics, 8*(1), 31–61.

Gries, St.Th. (2006). Corpus-based methods and cognitive semantics: The many senses of *to run*. In Gries, Stefan Th. & Anatol Stefanowitsch (Eds), *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis* (pp. 57–99). Berlin, Heidelberg, New York: Mouton de Gruyter.

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

Gries, St.Th., & Divjak, D. (In press). Behavioral profiles: a corpus-based approach to cognitive semantic analysis. In V. Evans & S. Pourcel (Eds.), *New Directions in Cognitive Linguistics.* Amsterdam and Philadelphia: John Benjamins.

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

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London and New York: Routledge.

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

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

Lakoff, G. (1990). The invariance hypothesis: Is abstract reason based on image-schemas? *Cognitive Linguistics, 1*(1), 39–74.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar: Theoretical Prerequisites.* Stanford: Stanford University Press.

Mondry, H., & Taylor, J. R. (1992). On lying in Russian. *Language and Communication, 12*(2), 133–43.

Pierrehumbert, J. (2001). Exemplar dynamics: word frequency, lenition, and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency and the Emergence of Linguistic Structure* (137–57). Amsterdam, Philadelphia: John Benjamins.

Quine, W. V. O. (Ed.) (1964). On what there is. In *From a logical point of view: nine logico-philosophical essays* (pp. 1–19). Cambridge, MA: Harvard University Press.

R Development Core Team. (2007). *R: A language and environment for statistical computing*. R Foundation for Statistical. Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http: // www.R-project.org.

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

Saaed, J. I. (1997). *Semantics*. Cambridge: Blackwell.

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.

Solovyev, V. D. (2006). Несколько замечаний о структуре семантического поля глаголов типа «стараться». [*Some remarks about the structure of the semantic field of verbs meaning 'try'*]. Unpublished manuscript, Kazan' State University.

Solovyev, V. D., & Bajraševa, V. P. (2007). О структуре семантического поля глаголов типа "стараться". [*On the structure of the semantic field of verbs like "starat'sja"*]. *Voprosy kognitivnoj lingvistiki 2*, 87–94.

Stefanowitsch, A., & Gries, St. Th. (Eds.). (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.

Wierzbicka, A. (1996). *Semantics: Primes and Universals*. Oxford: Oxford University Press.

Corresponding address:

Dagmar Divjak
Department of Russian and Slavonic Studies
Arts Tower, Western Bank
Sheffield
S10 2TN, UK.
Phone +44 114 222 7401, fax +44 114 2227416

E-mail: d.divjak@sheffield.ac.uk