Clusters in the Mind?

Divjak, Dagmar; Gries, Stefan

DOI:
10.1075/ml.3.2.03div

Clusters in the mind? Converging Evidence from Near-Synonymy in Russian.

Dagmar S. Divjak[1], Stefan Th. Gries[2]

[1]University of Sheffield

[2]University of California, Santa Barbara

Address for correspondence

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

Abstract

This paper provides experimental evidence to support the existence of mental correlates of lexical clusters. The results we present validate the linguistic model for nine near-synonymous verbs that express TRY-ing in Russian that was proposed on the basis of corpus data (Divjak and Gries 2006). 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. Consequently, the position of a corpus-based behavioral profile approach to lexical semantics is strengthened as it provides a firm basis for cognitively realistic language descriptions.

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

1. 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. Lakoffs (1990: 53) Cognitive Commitment). 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, while 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 of cognitive semantics did witness a gradual shift from intuition-based, corpus-illustrated work to corpus-based analyses (cf. Kishner and Gibbs 1996, Matlock 2001, and the papers in Gries and Stefanowitsch 2006 and Stefanowitsch and Gries 2006), few lexical semanticist (with the exceptions of Sandra and Rice 1995, Rice 1996, and Arppe and Järvikivi 2007) have taken on the challenge of validating relevant linguistic findings experimentally.

It comes as no surprise then that the above publications criticized the cognitive linguistic methodology, and in particular the widely used network representations of words and word senses, for a number of shortcomings. Among the most pressing questions are, no doubt: 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, say, polysemy or near-synonymy, make to issues of mental representation of lexical items?

In this paper, we will present results from research into near-synonymy in Russian that seeks to remedy these issues by relying on corpus data to construct a model for nine near-synonyms expressing TRY and validating the resulting model experimentally.

2. A Corpus-based Approach to Meaning

In recent work, Divjak and Gries have introduced what they refer to as the "behavioral profile"-approach, henceforth BPapproach , to lexical semantics (see Gries & Divjak 2008 for an overview). Given that the BP approach takes a usage-based view on meaning, and therefore we will use the words *use* and *meaning* interchangeably. Yet, although differences in usage can be of syntactic, semantic, pragmatic or socio-lectal nature, we will - with one exception - restrict our discussion to denotative aspects of meaning, thus leaving aside pragmatic and socio-lectal variation.

Since the BP approach is usage- based, , it qualifies asa data-driven and hence more objective means to capturing and comparing a word's meaning (Divjak 2006, Divjak and Gries 2006) or word senses (Gries 2006). In addition, behavioral profiles facilitate discovering the internal structure of polysemous or near-synonymous words as the profiles can be subjected to exploratory statistical techniques that find structure in large datasets, e.g. cluster analysis.

2.1 Tagging for meaning

We will illustrate the main characteristics of the BP approach using the study the results of which we seek to validate here: Divjak and Gries (2006) analyzed 1,585 sentences each containing one out of nine verbs that, in combination with an infinitive, express TRY in Russian, i.e. *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 availability) were annotated for 87 properties, a.k.a. levels of ID tags, listed in Table 1.

Table 1

Levels of ID tags used in annotating corpus extractions (adapted from Divjak and Gries 2006)

As a result, the 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 above listed. This dataset was analyzed using a hierarchical agglomerative cluster analysis, using the Canberra similarity metric and Ward's amalgamation strategy (for a more precise description of the procedure followed we refer to Gries and Divjak 2008). The resulting dendrogram, presented in Figure 1, shows what is similar and what is different: verbs that are clustered or amalgamated early are similar, whereas verbs that are amalgamated late are rather dissimilar. For example, it is obvious that *pytat'sja* and *starat'sja* are much more similar to each other than, say, *probovat'* and *norovit'*, which are only linked in the last overarching cluster. At the same time, the dendrogram gives an indication of how independent the clusters of verbs are: the larger the distance between different points of amalgamation, the more autonomous the earlier verb/cluster is from the verb/cluster with which it is merged later. In the present case, the plot clearly consists of three clusters and given the verbs and ID tag levels that were most strongly correlated with these clusters, Divjak and Gries (2006) labelled them YOU COULD SUCCEED, YOU WON'T SUCCEED and YOU CAN´T SUCCEED.

Figure 1

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

This dendrogram can easily be "translated" into a radial network so typical of cognitive linguistic analyses; this can be achieved either manually (Divjak 2004) or by means of phylogenetic clustering techniques (Divjak and Gries 2006). Yet, , BPs are not only an excellent basis for revealing the internal structure of a group of near-synonyms in a way compliant with fundamental cognitive linguistic assumptions: they also faciliate investigating 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, as attested in this dataset (cf. Backhaus et al. 1996:310-2). 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 particular situation displaying this property will be verbalized using a verb from that cluster. We will summarize the main findings of Divjak and Gries (2006) in the following section, yet given the large number of results yielded by this procedure, we restrict our attention to 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.

2.2 Evaluating the results

If we pull together the dimensions with the most revealing *t*-values[1] for the arguably most central and neutral YOU COULD SUCCEED cluster and incorporate them into one scenario, the characterization that emerges for *pytat'sja*, *starat'sja* and *probovat'*, is the following: 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 untimely brought to a halt (with *bylo*, *t*=0.982), that the subject was exhorted to undertake an attempt (*t*=0.832) and that the intensity with which the attempt was carried out was reduced (*t*=0.667).

In the YOU WON'T SUCCEED cluster with *silit'sja*, *poryvat'sja* and *norovit'*, an inanimate subject undertook repeated non-intense attempts to exercise physical motion; the actions are often uncontrollable and fail because of internal or external reasons. All three verbs lack a perfective counterpart and prefer the present tense more than verbs in the two other clusters (t=1.047 for present tense with a perfective infinitive and t=0.711 for the present tense followed by an imperfective infinitive). Among the most strongly represented variables we encounter the verbs' compatibility with inanimate subjects, both concrete and abstract (t ranges from 1.108 to 1.276), as well as with groups or institutions (t=1.297). Actions expressed by the infinitive are physical (t=0.176), affect a theme/patient (t=0.352), are metaphorical extensions of physical actions (t=0.999), or physical actions affecting a theme/patient (t=0.175). Focus is on the vainness (t=0.962 for vainness combined with intensity) of the durative effort (t=0.750 for duration adverbs).

With verbs from the YOU CAN´T SUCCEED 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 are metaphorical extensions of physical actions. 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.

There is one important disclaimer that applies to all of the above: analyses of corpus-data single out properties that are important within a particular dataset and are likely to generalize beyond a particular dataset. Yet, a radial network for near-synonyms expressing TRY constructed on the basis of a linguistic data analysis alone is by no means necessarily a truthful depiction of the representation present in the mind of speakers (cf. Sandra and Rice 1995). Put differently, while the usage-based view on language prominent within Cognitive Linguistics places quite some 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 methods.

3. Exploratory Analysis

There are indications that there is cognitive reality to the clustering obtained for nine near-synonymous verbs that epressTRYin Russian (see Figure 1): the results from a preliminary sorting task (Solov'ev, ms.) revealed that each of the nine verbs is most often grouped together with one of the verbs it is clustered togeher with in the corpus-based analysis. Yet, additional experiments and more refined evaluation techniques are needed to validate the findings; the results will be presented in Sections 4.2 and 4.3 respectively.

3.1 A first exploratory sorting task (Solvyev, ms.)

Solovyev (ms.) 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 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 then calculated how often subjects had grouped them together

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

However, in order to facilitate comparison of the experimental results with the corpus-based results as well as to homogenize the method of evaluation across different types of experiments (see below), we have designed an evaluative approach based on a point-scoring system that consists of two steps: first, we quantify the fit of our experimental results to the corpus results by means of a score; second, we compute a random baseline to assess whether the obtained fit could have been obtained on the basis of chance alone. In what follows, we explain our method evaluation in more detail.

3.2 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: *silit'sja, pryvat'sja,* and *norovit'*;
- cluster 2: *probovat', pytat'sja,* and *starat'sja*;
- cluster 3: *tscit'sja, pyzit'sja*, and *tuzit'sja*.

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

Table 2

Co-classification matrix (data from Solovyev, ms.)

This symmetric matrix has an unpopulated main diagonal since each verb *v* is by definition sorted into the same group as *v* itself. Second, in order to avoid basing our conclusions on raw frequencies of occurrence only, we compute each cell's Pearson residual (as it would result from the application of a chi-square test). Pearson residuals are obtained as shown in (1): positive versus negative values indicate that a particular frequency is higher or lower than expected by chance respectively.

(1) Pearson residual = $\frac{observed - \exp ected}{\sqrt{\exp ected}}$

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 3

Pearson residuals for the co-classification matrix in Table 2

Next, a point score needs to be computed that quantifies how well the sorting data fit the corpus data: since a high Pearson residual in Table 3 reflects that, say, *norovit'*, was sorted together with *poryvat'sja* much more often than expected by chance, we adopt 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, we 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, we scored zero points.

Since all verbs except for *silit'sja* have their highest Pearson residual for another verb from the same corpus-based cluster we score 8 points. However, it is yet unclear whether this score signals a good or a bad fit and whether or not this fit can be expected to occur by chance. We therefore test the fit for significance using a simulation-based approach.

From Table 3, it is clear that the minimum and the maximum scores that can be observed are 0 and 9 points respectively. It is also clear, then, that 8 points is a very good result. To test whether this result is sufficiently – i.e., significantly – different from chance, we first enumerated all scores any verb could possibly obtain. Since each verb is part of a three-verb cluster, this means that each verb 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. To test this result for significance and in order to avoid 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}, sample one value from this vector nine times (once 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 8 (our observed value) or higher: this turned out to happen 12 out of all 100,000 times; thus, $p_{one\text{-}tailed}$=0.00012, which shows that the observed value of 8 is not only approximately 3.5 times higher than expected by chance, but also highly significantly so. Table 4 contains the most important quantiles resulting from the simulation.

Table 4

Quantiles from the simulation

The results of Solovyev's (ms.) sorting experiment support the three cluster-solution that was arrived at on the basis of the corpus data. Admittedly, Solovyev (ms.), an as yet unpublished study, elicited sortings in a rather crude way, i.e. without providing the intended syntactic and semantic context for the verbs. In Section 4.1, we discuss the results from our own sorting-experiments that we followed up with a gap-filling task (Section 4.2).

4. Two Experiments

In this section we aim to provide an answer to two questions related to the degree to which the corpus-based results are corroborated by experimental evidence and the degree to which corpus-based studies contribute to linguistic investaigations of semantic and conceptual issues. Do native speakers produce groups that resemble the clustering obtained from analysis of corpus-data (or do they prefer the traditional pairs)? 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?[3] (Cf.

Arppe & Järvikivi 2007 for the synonym pair *miettiä* and *pohtia*)

Before embarking on the analysis, one caveat is in order. Whenever reference is made to the "cognitive reality of model", no position is taken as to the exact mental representation or mental storage of lexical clusters. Whichever way lexical information is stored, it is very well suited to produce clusters and it seems to include information about disinctive properties as they fall out from a corpus-driven linguistic analysis.

4.1 Three Sorting Tasks

4.1.1 Experimental design

46 third year IT students students from the Moscow Steel and Alloys Institute (www.misis.ru), Department of Computer Science and Economics[4] were presented with a questionnaire that contained instructions for three sorting tasks. In each task the subjects were presented with nine sentences that differed only with respect to the main verb expressing TRY that was used. The schematic sentence and its translation are given in (2); the underlined gap was filled by past tense forms of the nine verbs meaning TRY in Russian.

(2) a. После операции калека _____________ ходить без помощи костылей.

b. After the operation, the cripple tried to walk without the help of crutches.

In task 1, the subjects were asked to sort the nine sentences into a number of groups of their choice 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 that they thought belonged in 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 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 sorting, offering us different standards of comparison for our corpus-based resultsthis will be discussed in the following section.

4.1.2 Results

The data were evaluated in the same way as Solovyev's (ms.) 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 Table 5, Table 6, and Table 7 for task 1, task 2, and task 3 respectively.

Table 5

Pearson residuals for the co-classification matrix of task 1

Table 6

Pearson residuals for the co-classification matrix of task 2

Table 7

Pearson residuals for the co-classification matrix of task 3

The point score resulting from each of these tables is 8: all verbs but *silit'sja* prefer to be grouped with verbs from the cluster they were associated with in the corpus-based clustering solution.

For each of these three results, we computed the same simulation as presented above for Solovyev's data. In all three cases, the results were identical: for all tasks, a point score of 8 or higher was obtained 12 times out of all 100,000 simulation runs; thus $p_{one\text{-}tailed}$=0.00012; consider also Table 8 for the quantiles of each task's simulation.

Table 8

Quantiles from the simulation task 1, task 2, and task 3

Thus, we find that the subjects – regardless of the exact sorting instructions they were given – strongly prefer sorting solutions that corroborate the corpus-based clustering: throughout, the point scores obtained are 3.5 times as high as expected by chance and that ratio difference is highly significant according to three Monte Carlo simulations with 100,000 runs each. 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 out of 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.[5]

Additional confirmation for the existence of three clusters in the elicited data that strongly resemble those found in the corpus-data comes from computing cluster analyses on each of the co-classification matrices from task 1 through task 3. We computed a hierarchical agglomerative cluster analysis on the co-classification matrix of each task, and in order to rule out methodological artifacts, we applied the same settings as Divjak and Gries (2006) applied to their corpus data (similarity measure: Canberra, amalgamation rule: Ward); the resulting dendrograms are shown in the appendix, together with some comments on the number of clusters and the quality of the clustering solutions. As is obvious, in each task all verbs but *silit'sja* end up in the same cluster as in the corpus data; 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, for the fit of each clustering of one of the sorting tasks and the corpus-based clustering of Section 1 and achieved the high value of 0.74 for each case.

4.2 A Gap-filling Task

4.2.1 Experimental design

In addition to the above sorting experiment, we performed a gap-filling experiment (similar to the one employed by Dąbrowska, to appear) to check whether there was a quantitative dimension to the ID tag levels that had been singled out as highly distinctive for clusters using *t*-scores. Arguably, the *t*-values resulting from cluster analysis are 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 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. Yet, cue validity is based more directly on probability than *t*-values: are *t*-values are linked to distribution, with tail values being less likely.

Again, subjects were presented with a questionnaire containing a list of 27 verbs (each of the nine verbs three times) as well as 27 sentences. The 27 sentences were taken from the Russian dataset on which the corpus analysis was based: for each of the nine verbs, we took three sentences that exhibited particularly *high t*-values for the verb in question and deleted the main verb from the sentences. A detailed enumeration of these properties was provided in section 2.2 and we will limit ourselves here to summarizing the ID tags used per cluster.

The cluster [YOU COULD SUCCEED] 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 [YOU CAN'T SUCCEED] 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, [YOU WON'T SUCCEED] 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 in-/external reasons. These three TRY verbs are often found as participles.

The questionnaires were presented to 45 students from a technical university in Moscow; they were asked to fill the gaps with the verbs from the list.[6]

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

4.2.2 Results

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 stimulus sentence from which the verb meaning 'try' that was used in the corpus example had been deleted and were asked to enter that of the nine verbs they considered most fitting. By analogy to the above procedure, we therefore begin 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 8 provides this gap-filling preference matrix.

Table 8

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 being used as a gap-filler. Second, we computed each cell's Pearson residual in the same way as above and provide all Pearson residuals for Table 8 in Table 9.

Table 9

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

The third step again consists of computing a point score that quantifies how well the corpus data fit the gap-filling preferences, but this time there is a slight change. Again, a high Pearson residual in Table 9 reflects that one verb was much more often provided as a gap-filler for another verb, but this time, there is a third scoring option, namely the possibility that the deleted stimulus verb is 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.

As before, the bold-typed figures in Table 9 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 sufficiently – i.e., 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 if it most strongly preferred itself as a gap-filler;

– 1 for either of two verbs from the same cluster;

– 0 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 this for significance, we therefore generate a vector with all possible scores {2, 1, 1, 0, 0, 0, 0, 0, 0}, and sample with replacement nine values from this vector (one for each verb), and add 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 turned out to happen 251 out of all 100,000 times; thus, $p_{one\text{-}tailed}$=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

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, which 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.

The results from comparing the cluster trees are not quite as supportive: 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 1 is 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 since each stimulus sentence only differed with respect tothe 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. Yet, since we wanted to chose authentic sentences, each sentence also contains a variety of additional *t*-values; this results in (weak) associations to (verbs from) other clusters. Thus, while the *t*-values according to which we selected the stimuli does result in the hypothesized gap-filling patterns (on the whole), the results for the gap-filling experiment are not as pronounced as those for the sorting data.

5. Conclusions

Clusters "exist" in corpus and mind. Our findings reveal that the corpus-based model we proposed (Divjak and Gries 2006) is not a by-product of corpus composition or of a statistical technique used, i.e. cluster analysis will always output structure; instead there seems to be a mental reality corresponding to clusters of near-synonyms. Our study thus yields relevant findings on all three levels of cognitive semantic analysis, i.e. the descriptive, methodological and theoretical levels.

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 that 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 near-synonyms into clusters, not pairs. Hence, near-synonymy is (at least) a graded triadic phenomenon: it 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.

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 have knowledge of the overall similarities between the nine near synonyms: our subjects sorted the nine near-synonyms into groups that correspond to the corpus-derived clusters and intersubstitutability between verbs from different semantic clusters proved to be rather low. Subjects are also 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 basis of corpus-research.

Remains the question of how the match arises between the corpus-based distributional findings and the experimentally-observed preferences. In our view, our results provide additional support for the hypothesis put forward by Dąbrowksa (to appear)[7]: 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 phonologically and semantically more abstract generalizations, or schemas, on the basis of recurrent exposure to particular components of meaning. In other words, with Abbot-Smith and Tomasello (2006:275) we argue for a 'hybrid' usage-based 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 a cloud of exemplars in what one might want to call, for the lack of a better term, "syntactic-semantic space". Whenever a speaker encounters yet another instance of one of the nine verbs meaning TRY, the memory representation of these verbs and their concrete uses is updated with the information contained in the most recent usage event. However, not all actual instances will be remembered: memory traces decay over time and while particular salient usage events may remain accessible, what remains for the most part are generalizations based on many similar but now forgotten usage events. These generalizations 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.

The results of the sorting and the gap-filling task then 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 accessing 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 arrangements of verbs in syntactico-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:

- the BP approach 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, Divjak and Gries 2006, and Dąbrowska to appear) 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);
- it correlates strongly with different experimental methods: sorting and gap-filling (cf. above and Dąbrowska to appear), sentence elicitation and video descriptions (cf. again Dąbrowska to appear).

We therefore hope that, as more and also more diverse corpora become available, this method of investigation will be more frequently applied within cognitive lexical semantics.

Appendix

In this appendix, we present and the results from the cluster analyses of the three tasks of our sorting experiment.

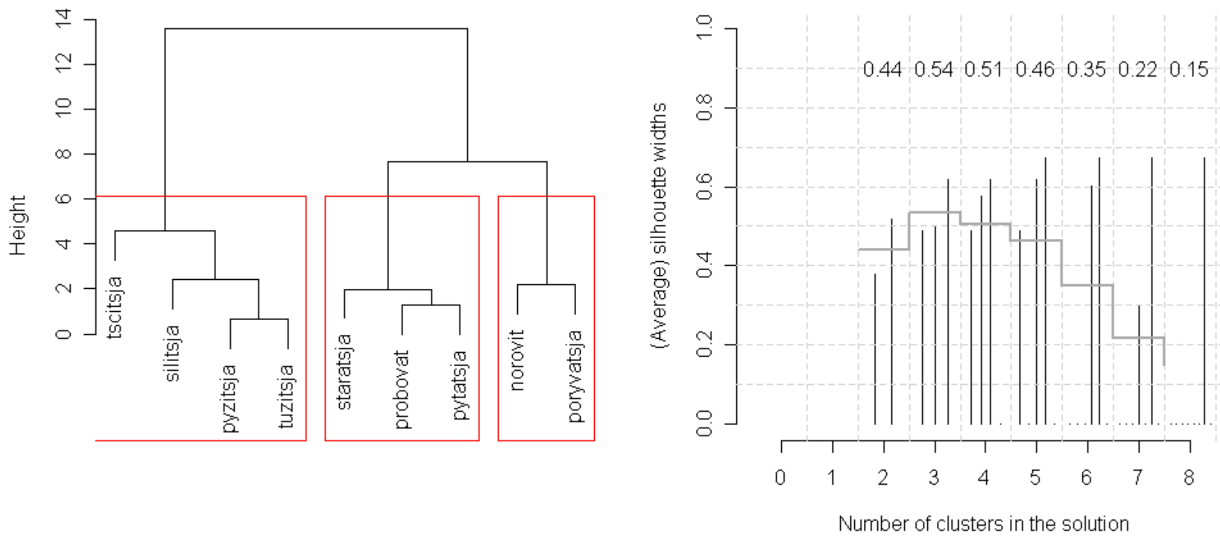


Figure (i): Cluster analysis for task 1 of our sorting experiment

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.

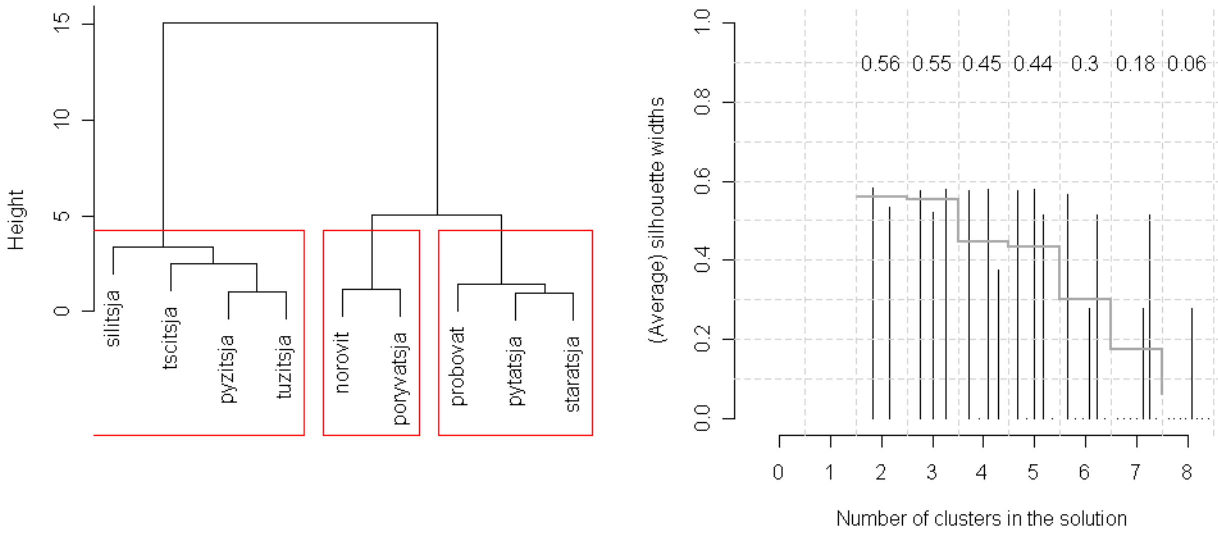


Figure (ii): Cluster analysis for task 2 of our sorting experiment

Here, both 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 are, therefore, compatible with the corpus-based solution.

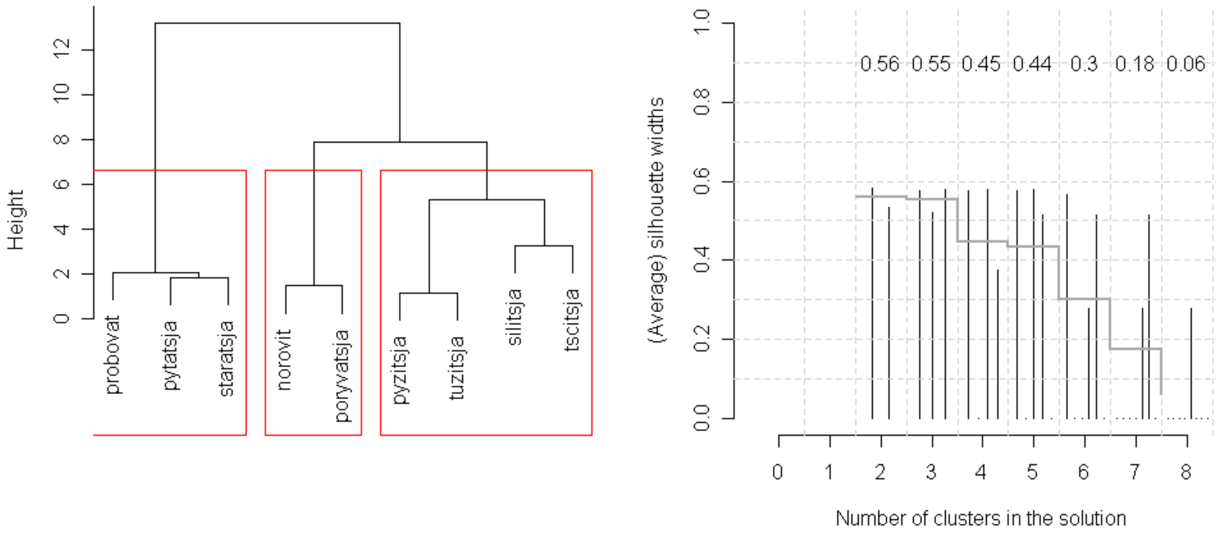


Figure (iii): Cluster analysis for task 3 of our sorting experiment

The results of this cluster analysis are interpreted as discussed under the clustering for task 2.

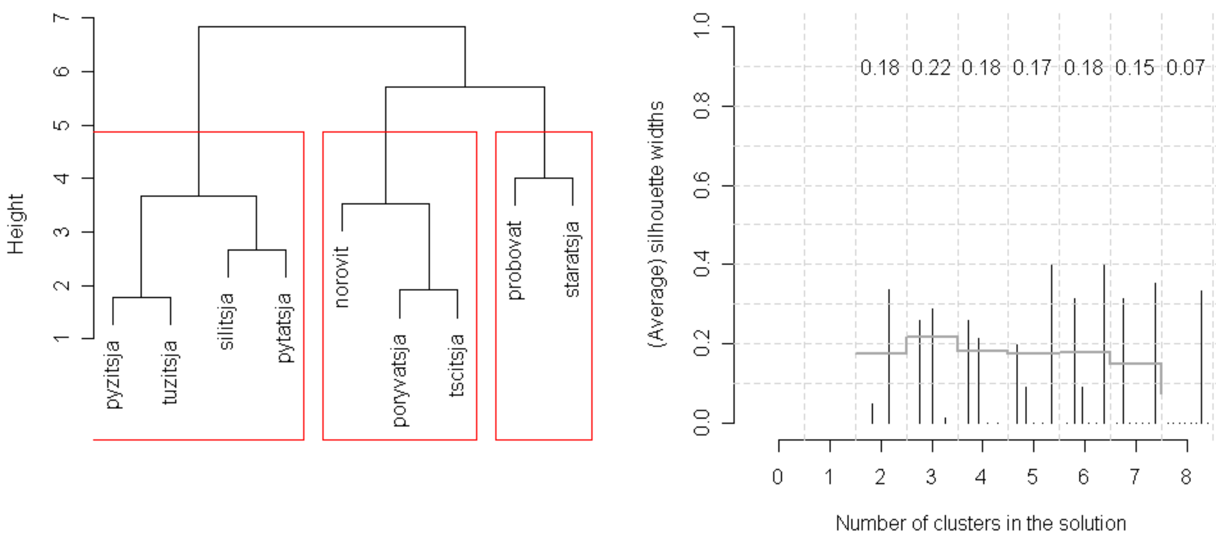


Figure (iv): Cluster analysis for our gap-filling experiment

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.

Notes

1 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 per 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 group, let alone to the same cluster of verbs.

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

3 We thank Leonid Oknyansky for assisting us in constructing the experimental materials.

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

5 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.

6 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.

7 Dąbrowska (to appear) 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.

Table 1

ID tags used in annotating corpus extractions (adapted from Divjak and Gries 2006)

Kind of ID tag	ID tag	Levels of ID tag
morphological	tense	present, past, future
	mode	infinitive, indicative, subjunctive, imperative, participle, gerund
	aspect	imperfective vs. perfective
syntactic	sentence type	declarative, exclamative, imperative, interrogative
	clause type	main vs. dependent
	type of dependent clause	adverbial, appositive, relative, zero-relative, zero-subordinator, etc.
semantic	semantic types of subjects, objects, etc.	concrete vs. abstract, animate (human, animal) vs. inanimate (event, phenomenon of nature, body part, organization/institution, speech/text) etc.
	properties of the process denoted by the verb	physical actions, perception, communication, intellectual activities, emotions, wishes/desires etc.
	controllability of actions	high vs. medium vs. no controllability
	adverbs, particles, connectors	temporal, locative, etc.
	negation	present vs. absent, attached to which element

Table 2

Co-classification matrix (data from Solovyev, ms.)

	norovit'	*poryvat'sja*	*silit'sja*	*probovat'*	*pytat'sja*	*starat'sja*	*pyzit'sja*	*tscit'sja*	*tuzit'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
pyzit'sja	1	2	20	0	2	4		7	27
tscit'sja	2	0	5	0	1	1	7		5
tuzit'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*	*pyzit'sja*	*tscit'sja*	*tuzit'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
pyzit'sja	-2.05	-1.46	3.39	-2.51	-2.01	-0.96		2.49	**5.5**
tscit'sja	-0.06	-1.36	0.95	-1.47	-0.99	-0.82	**2.49**		1.15
tuzit'sja	-1.36	-1.98	3.4	-2.21	-2.13	0.05	**5.5**	1.15	

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

Table 5

Pearson residuals for the co-classification matrix of task 1

	norovit'	*poryvat'sja*	*silit'sja*	*probovat'*	*pytat'sja*	*starat'sja*	*pyzit'sja*	*tscit'sja*	*tuzit'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
pyzit'sja	-2.56	-3.04	1.73	-2.93	-3.26	-2.32		0.19	**4.39**
tscit'sja	-0.75	-1.59	-0.15	-2.9	-2.97	-2.73	0.19		**0.36**
tuzit'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*	*pyzit'sja*	*tscit'sja*	*tuzit'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
pyzit'sja	-2.25	-2.83	0.58	-3.04	-3.24	-2.92		3.05	**4.22**
tscit'sja	-2.54	-1.87	0.07	-3.38	-2.8	-2.97	**3.05**		1.96
tuzit'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*	*pyzit'sja*	*tscit'sja*	*tuzit'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
pyzit'sja	-2.2	-2.88	1.69	-3.68	-3.16	-2.95		0.45	**5.01**
tscit'sja	-0.12	1.28	-2.09	-3.14	-2.61	-3.43	0.45		**1.76**
tuzit'sja	-2.68	-2.86	1.47	-3.67	-3.39	-3.44	**5.01**	1.76	

Table 8

Quantiles from the simulation task 1, task 2, and task 3

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

Table 9

Gap-filling preference matrix

	Response								
Stimulus	*norovit'*	*poryvat'sja*	*silit'sja*	*probovat'*	*pytat'sja*	*starat'sja*	*pyzit'sja*	*tscit'sja*	*tuzit'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
pyzit'sja	7	4	8	5	20	22	35	20	13
tscit'sja	19	21	22	6	18	20	3	13	11
tuzit'sja	12	5	20	3	20	14	17	27	13

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*	*pyzit'sja*	*tscit'sja*	*tuzit'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
pyzit'sja	-2.14	-2.94	-1.82	-2.53	1.08	1.62	**5.94**	1.38	-0.24
tscit'sja	0.95	1.4	**1.83**	-2.25	0.6	1.14	-2.81	-0.42	-0.75
tuzit'sja	-0.78	-2.63	1.38	-3	1.18	-0.32	1.1	**3.33**	-0.16

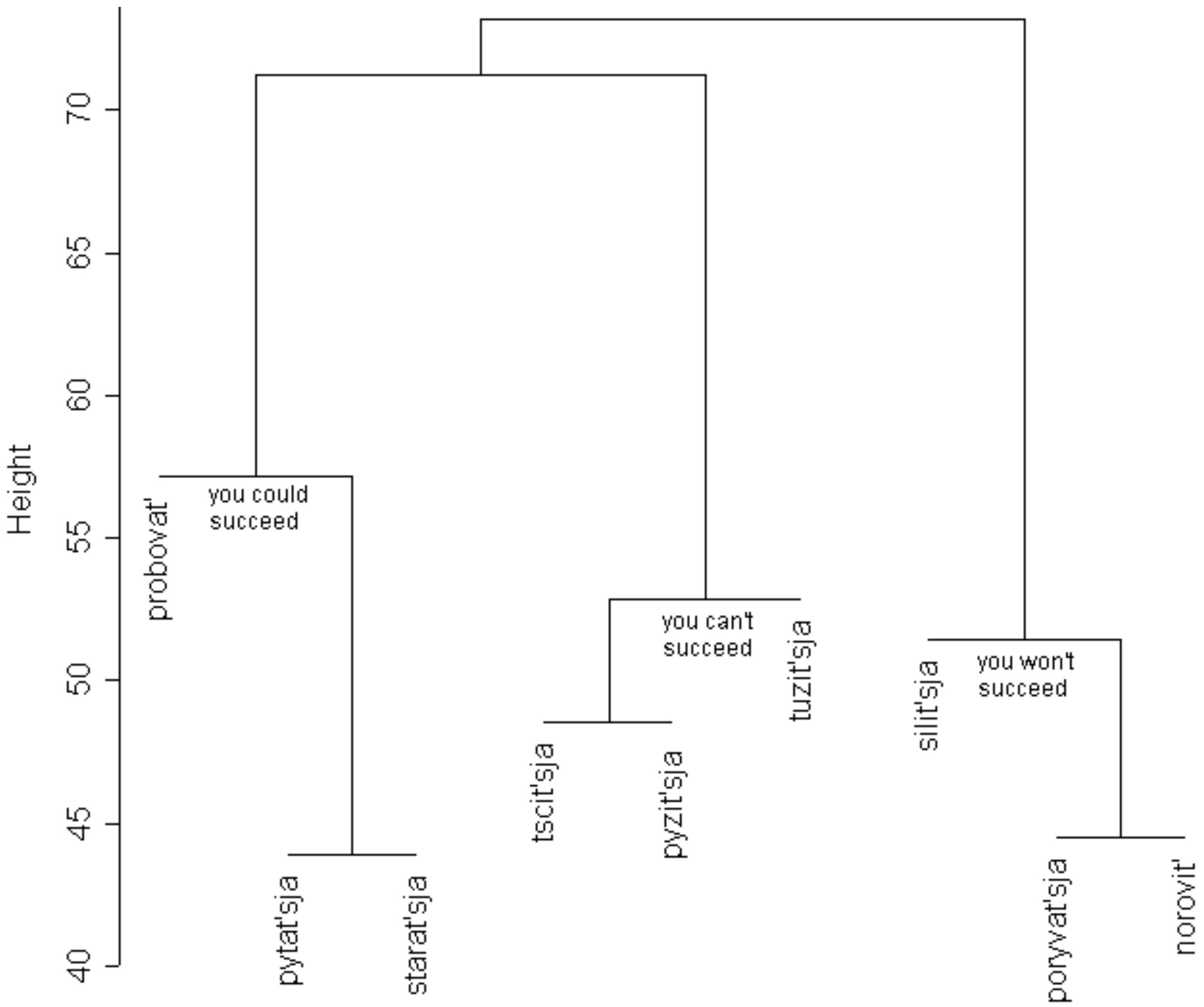
Table 11

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 1

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



References

Abboth-Smith, Kirsten and Michael Tomasell0. 2006. Exemplar-learning and schematization in a usage-based account of syntactic acquisition. The Linguistic Review 23, 275–290

Dąbrowska, Ewa. To appear. Words as constructions. In: Evans, Vyvyan and Stéphanie Pourcel (eds.). *New Directions in Cognitive Linguistics.* Amsterdam and Philadelphia: John Benjamins.

Divjak, Dagmar S. 2004. *Degrees of Verb Integration. Conceptualizing and Categorizing Events in Russian*. Unpublished Ph.D. dissertation, Dept. Oriental and Slavic Studies, K.U.Leuven, Belgium.

Divjak, Dagmar S. 2006. Ways of intending: A corpus-based cognitive linguistic approach to near synonyms in Russian. In: Gries, Stefan Th. and Anatol Stefanowitsch. (eds.). *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis.* Berlin, Heidelberg, New York: Mouton de Gruyter, 19-56.

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

Fowlkes, E.B. and C.L. Mallows. 1983. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association* 78.383:553-69.

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

Gries, Stefan Th. and Dagmar S. Divjak. To appear. Behavioral profiles: a corpus-based approach to cognitive semantic analysis. In: Evans, Vyvyan and Stéphanie Pourcel (eds.).

New Directions in Cognitive Linguistics. Amsterdam and Philadelphia: John Benjamins.

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

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.

Solov'ëv, Valerij. unpubl. ms. Несколько замечаний о структуре семантического поля глаголов типа "стараться".