Does language-as-used fit a self-paced reading paradigm?

Divjak, Dagmar; Arppe, Antti; Baayen, Harald

Does language-as-used fit a self-paced reading paradigm? (The answer may well depend on how you model the data.)

Dagmar Divjak <d.divjak@sheffield.ac.uk> – corresponding author[1]
Antti Arppe <arppe@ualberta.ca>
Harald Baayen <harald.baayen@uni-tuebingen.de>

Abstract

We report on a self-paced reading experiment that was run to ascertain whether the effect of differential tense, aspect and mood (henceforth TAM) marking on verbs would affect processing. TAM properties were identified as the strongest predictors for the choice between 6 near synonyms meaning TRY in Russian on the basis of regression models fit to manually annotated corpus data (Divjak 2010, Divjak & Arppe 2013). We will discuss how we used a Generalized Linear Mixed Model to account for the fact that we deviated from the traditional set-up for self-paced reading in two ways: we used an imbalanced design and ran the task with actually attested sentences rather than artificially created ones. These deviations were motivated by the need to accommodate the natural restrictions on TAM combinations and to respect the lack of a strict word order, which are both typical for Russian. We will also describe how we used a Generalized Additive Model to handle the non-linearities that we encountered in the reading times data.

[1] Author contributions: DD and AA conceived and designed the self-paced reading experiment; DD ran the experiment; DD, AA and HB analyzed the data with comments from Petar Milin; DD wrote the paper using comments and suggestions from AA and HB. The PsychoPy script for self-paced reading was written by Lily FitzGibbon; participants were recruited and scheduled by Daria Satyukova. The experiment received ethical approval from the University of Sheffield, School of Languages & Cultures. The financial support of the Prokhorov Foundation and the logistic support of the Saint Petersburg branch of the Russian Academy of Sciences are gratefully acknowledged.

0. Introduction: From text to model to mind?

Over the past 15 years probabilistic statistical classification models have become established as de facto methodological standard for predicting the choice between lexical or constructional alternatives in usage-based linguistics. It is a method widely applied in semantics (e.g. Arppe & Järvikivi 2007, Arppe 2008, Divjak 2010, Divjak & Arppe 2013), syntax (e.g. Gries 2003, Bresnan 2007, Bresnan et al. 2007, Bresnan & Ford 2010, Kendall et al. 2011, Klavan 2012), morphology (e.g. Antić 2012, Baayen et al. 2013), phonetics and phonology (e.g. Erker & Guy 2012, Raymond & Brown 2012) and in areas as diverse as sociolinguistics (e.g. Grondelaers & Speelman 2007), historical linguistics (e.g. Gries & Hilpert 2010, Szmrecsanyi 2013, Wolk et al. 2013) and language acquisition (e.g. Ambridge et al. 2012).
We are currently experiencing another shift (see Klavan & Divjak 2016), i.e. towards providing experimental "validation" for such models (cf. a series of papers by Arppe & Järvikivi 2007, Bresnan & Ford published between 2007 and 2013, Divjak et al. 2016, Klavan 2012, Perek 2015). Bearing in mind the age-old adage that "[n]ot everything that counts can be counted and not everything that can be counted counts" we indeed need to ask the question of what is real about such statistical classification models. There are two aspects to this question. On the one hand, it addresses one of the main problems that corpus-linguists face when annotating datasets for their research, i.e. the decision on the level of granularity: which level of annotation and which annotation scheme yield the best prediction. On the other hand, it targets a key concern for cognitive corpus linguists: there is an abundance of patterns that can be detected in usage, but what the analyst detects may well be different from what the speaker detects and uses. Is the model that we propose cognitively realistic? Can we by means of textual data analysis get at what drives speakers?

In this chapter, we focus on the choice between near-synonymous verbs expressing TRY in Russian that, among an impressive list of other synonymous words, have been the subject of extensive study by linguists from the Moscow Semantic School. In 20th century Western Linguistics, on the contrary, synonymy was rather neglected: part of the reason for this might be that a graded, lexical phenomenon like near-synonymy does not fit in well with the theoretical frameworks that predominated Western linguistics during the second half of the 20th century. During that time, synonymy was reserved for lexicographers, who often worked in a corpus-illustrated fashion (Tognini-Bonelli 2001). Early studies focused on pairs of synonyms, e.g. Geeraerts (1985) on *vernietigen* and *vernielen* (destruct/destroy) in Dutch, Church et al. (1991; 1994) on *strong* versus *powerful*, Mondry & Taylor (1992) on lying in Russian (*lgat'* versus *vrat'*), Schmid (1993) on *start* versus *begin,* Taylor (2003) on *high* versus *tall*, Kjellmer (2003) on *almost* and *nearly*. Biber et al. (1998) studied a group of 3 synonyms: *big*, *large* and *great*, while Gries (2003) compared similar adjectives ending in *–ic* and *-ical*. Divjak (2004, 2010) attempted to put the study of lexical synonymy on sounder footing, thereby testing assumptions from usage-based theory in general and from cognitive linguistics in particular. At the same time, Arppe (2008) approached synonymy from a theoretically agnostic, quantitative perspective, while a primarily computational linguistic approach is presented in long-standing, comprehensive work by Hirst and collaborators (e.g. Edmonds & Hirst 2002; Inkpen & Hirst 2006).

After an introduction to synonymy (Section 1), we review the corpus-based analysis of TRY verbs in Russian (Section 2) before moving on to new data from a self-paced reading task that was run to assess whether the effect of the strongest predictors would be felt during processing in Section 3. In Section 4 we reflect on the linguistic insights that were gleaned from working on data from a morphologically rich language.

1. What is synonymy?

Traditionally, two words are considered synonymous in a sentence or linguistic context if the substitution of one for the other does not alter the truth value of the sentence. Two lexical units would be absolute synonyms if and only if all their contextual relations were identical. For this reason, it is commonly asserted that absolute, perfect or full synonyms do not exist. Synonyms, then, are defined as lexical items whose senses are identical in respect of "central" semantic traits, but differ in respect of so-called "minor" or "peripheral" traits.
Within the Western tradition (Cruse 2000), synonyms are defined contextually by means of diagnostic frames. For cognitive synonyms such as *die*, *pass away* and *kick the bucket* that only differ in expressive traits it is impossible to state **He kicked the bucket but he did not die*. Yet plesionyms differ in more than just expressive traits, so two plesionyms can be united in one sentence such as *He was killed, but I can assure you he was not murdered*. In the Russian tradition, the decompositional approach prevails and synonyms are analyzed by means of a semantic metalanguage. Apresjan et al. (1995: 60, 2000: XL) defines the constitutive characteristic of "synonyms" as "the presence in their meaning of a sufficiently big overlapping part". To define the "sufficiency" of "big overlapping", the meanings of words are reformulated with the help of a special meta-language. The strict formulation prescriptions and the limited inventory of lexical primitives of this metalanguage facilitate comparison of meanings. The overlap has to be bigger than the sum of the differences for two lexemes, or at least equal to the sum of the differences in case of three or more lexemes. Apart from that, the overlap has to relate to the assertion of the definition that contains "genera proxima", the syntactic main word of which coincides.
On both accounts (for a detailed discussion of the pros and cons of these approaches and an alternative proposal see Divjak 2010), the three verbs that are in focus in this chapter (*пробовать, пытаться, стараться*) would qualify as near-synonyms and they constitute a separate entry in Apresjan et al. (1999), indeed. Yet, as explained in Divjak (2010: 1-14), the verbs were in fact selected on the basis of a distributional analysis in the tradition of Harris (1954) and Firth (1957), with meaning in the Wittgensteinian sense construed as contextual. Synonymy was operationalized as mutual substitutability (i.e., interchangeability), within a set of constructions, i.e. a shared constructional network. On a Construction Grammar approach to language both constructions and lexemes have meaning; as a consequence, the lexeme's meaning has to be compatible with the meaning of the construction in which it occurs and of the constructional slot it occupies to yield a felicitous combination. Therefore, the range of constructions a given verb is used in and the meaning of each of those constructions are revealing of the coarse-grained meaning contours of that verb. The results can then be used to delineate groups of near-synonymous verbs. On this approach, near-synonyms share constructional properties, even though the extent to which a construction is typical for a given verb may vary and the individual lexemes differ as to how they are used within the shared constructional frames.

2. Fitting a polytomous regression model to corpus data on TRY verbs in Russian

For an exhaustive overview of corpus-based work on Russian TRY verbs, we refer to Divjak (2010: 177-193). Here we will focus on the corpus research that inspired the hypothesis tested using self-paced reading.

We build on earlier work by Divjak (2004/2010), who constructed a database containing 1585 tokens for 9 Russian verbs that mean TRY if combined with an infinitive: *probovat', pytat'sja, starat'sja, silit'sja, norovit', poryvat'sja, pyžit'sja, tščit'sja, tužit'sja.* The last 3 occur, however, too infrequently to yield reliable estimates in a regression model and were therefore omitted. Source of the data were the Amsterdam Corpus that contains written literary texts, supplemented with data from the Russian National Corpus. About 250 extractions per verb were analysed in detail, except for *poryvat'sja* that is rare and for which only half that number of examples could be found. Samples of equal size were chosen because of two reasons: 1) interest was in the contextual properties that would favour the choice of one verb over another, and by fixing the sample size, frequency was controlled, 2) the difference in frequency of occurrence between these 9 verbs is so large that manually annotating a sample in which the verbs would be represented proportionally would be prohibitively expensive.

The extractions were manually annotated for a variety of morphological, semantic and syntactic properties, using the annotation scheme initially proposed in Divjak (2003, 2004) and later described under the name Behavioral Profiling (BP) in a number of publications (Divjak 2006, Divjak & Gries 2006). Divjak's (2003, 2004) BP bears resemblance to annotation schemata used in Gries (2006) and Arppe (2008), and the name can be traced back to Hanks (1996), whose profiles were, however, restricted to complementation patterns and semantic roles. BPs chart behaviour of X across N contexts (where context = "natural" unit of expression, i.e. sentence or clause) for a multitude of parameters (incl. grammatical information). The net is cast wide because it is not known what does (not) convey meaning.

The tagging scheme (for details see Divjak (2010: 119-129)) was built up incrementally and bottom-up, starting from the grammatical- and lexical-conceptual elements that were attested in the data. This scheme captures virtually all information provided at the clause (in case of complex sentences) or sentence level (for simplex sentences) by tagging morphological properties of the finite verb and the infinitive, syntactic properties of the sentences and semantic properties of the subject and infinitive as well as the optional elements. All annotation is "naïve" (different from the Bresnan dative studies), meaning that only such linguistic labels are used for which linguistically naïve native speakers can reasonably be expected to have a matching category. For example, we do not expect native speakers to be able to identify an inanimate subject or a past tense, but we do expect them to know whether something is alive or whether an event has already happened. There were a total of 14 multiple-category variables amounting to 87 distinct variable levels or contextual properties (listed in Table 1), and yielding a set of 137,895 manually coded data points.

Type of variable	Variable name	Variable level name
morphological	tense	future, present, past, not applicable
	mode	infinitive, indicative, imperative, participle, gerund, conditional
	aspect (of both finite and infinite verb)	imperfective vs. perfective
syntactic	subject structure	nominative to the tentative verb, nominative to the preceding verb, accusative to the preceding verb, dative to the preceding "personal" verb, dative to the preceding "impersonal" verb, dative to the tentative verb, the subject is the infinitive tentative verb, the infinitive tentative verb modifies a noun
	sentence type	declarative, interrogative, imperative, exclamation
	clause type	main clause, subordinate clause
semantic	semantic type of subject	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, physical involving another, physical exchange/transfer, physical motion, physical motion involving another, physical figurative, physical figurative involving another, figurative physical exchange/transfer, figurative physical motion, figurative physical motion involving another, perceptual, perceptual active, communication/interaction, mental, emotional
	controllability of the infinitive action	high vs. medium vs. no controllability
	adverbial specification	duration (*dolgo* 'long', *dolgoe vremja* 'a long time'…), durative repetition (*vsë* 'all (the time)', *vsë vremja* 'all the time'…), repetition (… *raz* '(…) times'), intensity (*očen*' 'very', *izo vsech sil* 'with all one's might'…), vainness/futility (*zrja, naprasno, tščetno* 'in vain'…), intensity & vainness (*kak ni/ne* … 'however')
	particles	exhortation (*davaj* … 'let's, come on'), permission (*pust'* … 'let'), restriction (*tol'ko* … 'only, just'), permission & restriction (*pust' tol'ko* … 'let … only'), intensification (*daže* … 'even'), untimely halt (*bylo*)
	connectors	external opposition (*no, a, i ne*), internal opposition (*no, a, i ne*), introducing a *čtoby* 'in order to' clause, in a *čtoby* 'in order to' clause
	negation	present vs. absent; to the tentative verb, to the infinitive

Table (1): Variables used in the annotation of the corpus sample

Divjak and Arppe (2013) used this dataset to train a polytomous logistic regression model (Arppe 2008, 2013a, 2013b), predicting the choice for one of the verbs. As this model underlies the self-paced reading task that is the focus of this chapter, we will describe the regression modelling in some detail.

As a rule of thumb, the number of distinct variable combinations that allow for a reliable fitting of a (polytomous) logistic regression model should not exceed 1/10 of the least frequent outcome (Arppe 2008: 116). In this case, the least frequent verb occurs about 150 times, hence the maximum number of variable categories should be approximately 15. The selection strategy we adopted (out of many possible ones) was to retain variables with a broad dispersion among the 6 TRY verbs. This ensured focus on the interaction of variables in determining the expected probability in context rather than allowing individual distinctive variables, linked to only one of the verbs, to alone determine the choice. As selection criteria we required the overall frequency of the variable in the data to be at least 45 and to occur at least twice (i.e. not just a single chance occurrence) with all 6 TRY verbs. Additional

technical restrictions excluded one variable for each fully mutually complementary case (e.g. the aspect of verb form – if a verb form is imperfective it cannot at the same time be perfective and vice versa) as well as variables with a mutual pair-wise Uncertainty Co-Efficient *UC* value (a measure of nominal category association; Theil 1970) larger than 0.5 (i.e. one variable reduces more than ½ of the uncertainty concerning the other). Altogether 18 variable values were retained (11 semantic and 7 structural), belonging to 7 different variable types. These are listed in Table (2). The model specification thus by and large consists of TAM markings on the verbs and semantic properties of the infinitive.

	Property
1	declarative sentence
2	human agent
3	*try* verb in main clause
4	*try* verb in perfective aspect
5	*try* verb in indicative mood
6	*try* verb in gerund
7	*try* verb in past tense
8	subordinate verb in imperfective aspect
9	subordinate verb involves high control
10	infinitive designates an act of communication
11	infinitive designates an act of exchange
12	infinitive designates a physical action involving self
13	infinitive designates a physical action involving another participant
14	infinitive designates motion involving self
15	infinitive designates motion involving another participant
16	infinitive designates metaphorical motion
17	infinitive designates metaphorical exchange
18	infinitive designates metaphorical action involving another participant

Table (2): Predictors used by the Divjak and Arppe (2013) model

Using the values of these variables as calculated on the basis of the data in the sample, the model predicts the expected probability for each verb in each sentence. More interesting from a linguistic perspective, the model tells us how strongly each property individually is associated with each verb (e.g. *norovit'* and especially *poryvat'sja* are strongly preferred when the infinitive describes a motion event while *pytat'sja, starats'ja* and *silit'sja* are dispreferred in this context; *probovat'* does not have a preference one way or the other). This enables us to characterize each verb's preference(s) (Divjak 2010, Arppe & Divjak 2013, Arppe 2013b).

Assuming that the model "chooses" the verb with the highest predicted probability (though strictly speaking a logistic regression model is attempting to represent the proportions of possible alternative choices in the long run), its overall accuracy is 51.7% (50.3% when tested on unseen data) and resampling techniques confirm this. This is well above chance: since there are six verbs, chance performance would have been at 16.7%. Verb-wise model predictions are provided in Table (3): the highest values are on the diagonal, i.e. each verb is most often predicted as itself.

	norovit'	**poryvat'sja**	**probovat'**	**pytat'sja**	**silit'sja**	**starat'sja**	**[original]**
norovit'	**143**	32	4	36	17	18	**250**
poryvat'sja	22	**57**	1	19	8	12	**119**
probovat'	8	8	**189**	16	5	20	**246**
pytat'sja	44	21	47	**73**	35	27	**247**
silit'sja	23	22	0	30	**152**	14	**241**
starat'sja	34	13	45	26	45	**85**	**248**
[predicted]	**274**	**153**	**286**	**200**	**262**	**176**	**1351**

Table (3): Model accuracy

Table (4) summarizes the verb-specific odds per property for all six Russian verbs (details can be found in Divjak & Arppe 2013). Cells with a "+" signal significantly positive odds, i.e. in favour of the occurrence of a lexeme; cells with a "0" are neutral, i.e. do not favour or disfavour a specific verb, whereas cells with a "-" indicate odds for properties significantly against a lexeme.

Property/Verb	**Probovat'**	**Pytat'sja**	**Starat'sja**	**Silit'sja**	**Norovit'**	**Poryvat'sja**
(Intercept)	-	-	-	0	0	-
CLAUSE.MAIN	**+**	-	0	0	0	0
FINITE.ASPECT_PERFECTIVE	**+**	0	0	0	0	0
FINITE.MOOD_GERUND	-	0	**+**	**+**	-	0
FINITE.MOOD_INDICATIVE	+	0	0	0	0	0
FINITE.TENSE_PAST	0	**+**	-	**+**	-	+
INFINITIVE.ASPECT_IMPERFECTIVE	**+**	-	**+**	-	-	0
INFINITIVE.CONTROL_HIGH	0	**+**	**+**	+	**+**	**+**
INFINITIVE.SEM_COMMUNICATION	**+**	-	0	0	0	**+**
INFINITIVE.SEM_EXCHANGE	0	0	0	-	**+**	**+**
INFINITIVE.SEM_METAPH…_MOTION	0	0	0	-	**+**	0
INF….SEM_METAPH…_PHYS…_EXCH…	0	-	0	-	**+**	0
INF….SEM_METAPH…_PHYS…_OTHER	0	0	0	0	**+**	0
INFINITIVE.SEM_MOTION	0	-	-	-	**+**	**+**
INFINITIVE.SEM_MOTION_OTHER	0	0	-	0	**+**	**+**
INFINITIVE.SEM_PHYSICAL	**+**	**-**	0	0	**+**	0

INFINITIVE.SEM_PHYSICAL_OTHER	**+**	+	-	-	**+**	**+**
SENTENCE.DECLARATIVE	-	0	**+**	0	0	0
SUBJECT.SEM_ANIMATE_HUMAN	0	0	**+**	0	-	**+**

Table (4): verb specific odds per property for all six Russian verbs

Overall, infinitival semantics play a significant role for *norovit'* and to a lesser extent for *poryvat'sja*, but are much less relevant for the other four verbs: they play hardly any role for *probovat'* and *starats'ja* and some seem to be repelling *pytat'sja* and *silit'sja*. If we take a specific property, such as main clause (CLAUSE.MAIN), we see that it has significant positive odds in favor of *probovat'*, neutral ones for *silit'sja*, *starat'sja*, *norovit'* and *poryvat'sja*, and significant odds against *pytat'sja*. Moreover, the comparatively high odds of having a perfective finite verb (FINITE.ASPECT_PERFECTIVE) in favor of *probovat'* may stand out — this is due to the fact that *probovat'* is one of only three verbs that have a perfective counterpart, and the verb that occurs most frequently in the perfective aspect in the data.

Despite the relatively good fit we achieved, we have to face the inconvenient truth that "[w]henever we make a model [...], we are trying to force the ugly stepsister's foot into Cinderella's pretty glass slipper. It doesn't fit without cutting off some essential parts." (Derman 2011). This realization has prompted a series of experimental studies that were designed to compare different aspects of the corpus analysis to different types of human behaviour.

3**. Testing the predictions of the corpus-based model experimentally**

Without going into detail we can say that, overall, the corpus-based models did well and mimicked subjects' behavior on a range of tasks so "there must be something to them".[2] Yet, the fact that the "resulting" states in model and speaker yield comparable results does not imply that they were arrived at by (exactly) the same means: the properties that play a role in capturing off-line knowledge of a phenomenon need not be the same as those guiding on-line processing of that same phenomenon. For this reason, we set out to capture time-bounded effects on sentence processing tasks such as reading (cf. Bresnan & Ford 2010) – are the effects of the corpus-based predictors that seem to play a role in off-line studies also active on-line, while processing language?

Regression models fit to corpus data (Divjak 2010, Divjak & Arppe 2013; summarized in Section 2) show that Tense, Aspect and Mood (TAM) markers, often overlooked in lexical semantic studies, are strong predictors of choice when faced with 6 near-synonymous verbs expressing TRY. Different from semantic properties, which seem to define 3 out of 6 verbs rather well (*norovit'*, *poryvat'sja* and *probovat'*), the 3 more frequent verbs (*probovat'*, *pytat'sja*, *starat'sja*, but also *silit'sja*) are defined by a combination of preferred and dispreferred TAM markers, as Table (4) above shows.
The effect of TAM marking came out as stronger for predicting the choice of TRY verb than semantic properties of subject and infinitive action: using just TAM predictors vs. non-TAM

[2] See Divjak & Gries 2008 for gap-filling and sorting data on the clustered lexical model proposed in Divjak 2003 and Divjak & Gries 2006 and see Divjak et al. 2016 for forced choice and acceptability ratings data on the regression models described in Divjak 2010 and Divjak & Arppe 2013.

predictors from the original model with 6 verbs, MacFadden's pseudo R_L^2 (the relative reduction in Deviance (based on Log-Likelihood) gained by the model, in comparison to a null model) is substantially better for a model with TAM predictors at 0.219 vs. 0.129 for a model without, and the same applies for accuracy with 0.429 for a model with TAM predictors vs. 0.363 for a model without.

That TAM marking would be important is at the same time surprising and expected: TAM marking is not typically used to tell synonyms apart, but it is a reliable predictor as TAM marking is obligatory: the presence of TAM markers on every verb form increases the frequency with which these properties are encountered. This is especially likely in Russian and other morphologically rich languages, for which it may be (more) cognitively unrealistic (than for morphologically poorer languages) to track words at the lexeme level rather than at the inflected/declined level. Sinclair (2001) advances the argument that collocations are also active at the word-form level, not so much only at the lemma-level, and may indeed differ for various forms of the same lemma. Newman (2008) discusses support for low-level generalizations (studying linguistic behavior at the inflected level of words, as opposed to generalizing linguistic behavior at the lemma level) in corpus-based research from language acquisition research (not all word forms are acquired simultaneously), grammaticalization studies (grammaticalization can affect particular inflected forms only, e.g. the use of *going to* as progressive marker in English) and stylistics (where inflectional differences are typical for different genres). Psycholinguistic experimentation has confirmed that not all inflected forms of a lemma are associated with one and the same reaction time (Baayen et al. (1997) report storage for high-frequency noun plurals; Kostić and Havelka (2002) discuss different reaction times for different person and number forms of Serbian verbs in the future tense; Kostić and Mirković (2002) discuss the impact of inflectional forms of Serbian noun paradigms on reaction times).

3.1 Self-paced reading

To explore whether the factors identified on the basis of corpus analysis also play an active role in processing, and in particular whether an in lexical semantics rather neglected formal variable such as TAM deserves more attention, we ran a self-paced reading task. In the self-paced reading task, participants are presented with a sentence one word at a time on a computer screen and must press a button as quickly as possible each time they read a word; the exact timings of the button-presses is recorded. An example stimulus is given in (1):

(1)

И	не	пробуйте	понимать		
I	ne	probujte	ponimat’		
And	not	try IMPF IMPER PL	understand IMPF INF		
чужого	счастья	—	не	поймете	
čužogo	sčast’ja	—	ne	pojmete	
another’s GEN SG	happiness GEN SG	—	not	understand PF IND NON-PAST 2PL	

Don’t try to understand some else’s happiness — you can’t.

The dominant interpretation (Kaiser 2013) of what reading times reflect would have us expect that the verb with more probable TAM marking would require fewer resources

during reading, so that processing complexity during reading would decrease on predicted high-probability TAM markings for a specific verb, resulting in quicker reading speeds. On this interpretation, we expect to find a negative correlation between probability of occurrence and reading times for TAM combinations, with more typical TAM markings leading to quicker reading times because they require less processing. However, there is an alternative interpretation of reading times which attributes a slowdown in reading speed to a sudden drop in parsing uncertainty (Hale 2003, Levy 2008).

Participants

We recruited 39 (17 male, 22 female) adult native speakers of Russian, aged between 18 and 31 (mean 23.6, s.d. 3.3) and currently living in St. Petersburg. The subjects were not linguists, philologists or language students and except for one subject, had never before participated in a (psycho-)linguistic experiment of any kind.

Materials

The 3 verbs used, *probovat'*, *pytat'sja* and *starat'sja*, are the most frequent and neutral ones. Of all TRY verbs, these three are the most similar to each other (Apresjan 1999, Divjak 2003, Divjak & Gries 2006) so the differences between the verbs are very small. Preceding corpus research and experimental validation had provided a rich knowledge base and this was used when selecting stimuli in which there were no known confounds. The following procedure was followed to select stimuli:

1. A full polytomous logistic regression model was run for the 3 verbs of interest, *probovat'*, *pytat'sja* and *starat'sja*
2. We checked whether certain types of subjects or infinitives increased the preference for one of the 3 verbs and if they did, these subjects or infinitives were avoided in the experimental items. We found that
    a. physical activities increase the chances of *probovat'* being chosen
    b. mental activity, metaphorical motion activity, motion activity involving another participant, physical action involving another participant reduce the chances of *starat'sja* being chosen
    c. there was no effect of subject on any of the 3 verbs – all three verbs were neutral towards being combined with human animate subjects
3. We selected experimental sentences the following way
    a. we ran a model with TAM-related variables for the TRY verb and one that included semantics for the infinitive: including infinitive semantics in the model gives us more precise information about the reading speed to expect since every sentence will include an infinitive. The infinitive semantics does not affect the probabilities significantly, but does tweak them; without the infinitive, all probabilities would be the same for one specific TAM combination.
    b. for each of the 9 existing TAM combinations we selected the top sentences in terms of probability estimates for all three verbs: we took those sentences that keep us closest to what the probabilities would be without us knowing what the infinitive is like, which controls for the effect of infinitive semantics.
        i. imperfective indicative past

   ii. imperfective indicative present
   iii. imperfective gerund present
   iv. imperfective indicative future
   v. perfective indicative past
   vi. perfective gerund past – **not attested in our database**
   vii. perfective indicative future
   viii. imperfective imperative
   ix. perfective imperative
4. The list of stimuli was compiled as follows:
   a. we selected 3 examples for each of the 3 verbs for all 8 TAM combinations that were attested in our dataset. Although 9 combinations exist, for the perfective past gerund no cases were attested in our data. This means that no contextual and hence no probability estimates were available and this TAM combination was excluded from the experiment. Some combinations did not have sufficiently many attestations in our data, and for those we consulted the RNC; in all, 18 RNC examples with the same contextual properties as specified by the model were added to the dataset while 54 stem from the annotated corpus sample.
   b. these examples were divided over 3 experimental sets: set 1 gets 1st examples; set 2 gets 2nd examples; set 3 gets 3rd examples. We ensured that the imperfective future and the infinitive semantics were evenly distributed over all three sets. A third of all sentences was followed by a yes/no question that the subject s had to answer.
   c. every participant was presented with 1 example for each verb-by-TAM combination. These examples were interspersed with 24 filler items containing verbs of perception. The set was preceded with 5 practice sentences and randomized automatically for each subject.

This set-up deviates from the traditional approach to self-paced reading experiments in three important ways:

1. we used an imbalanced design to accommodate the natural restrictions on TAM combinations – e.g. there are no present perfectives and none were created for the experiment.
2. we ran the task with actually attested sentences rather than artificially created ones. Because we used authentic examples, all stimuli were possible, albeit more or less likely.
3. working with authentic sentences also introduced variation in the position the TRY verb occupied in the sentence. This is exacerbated by the fact that Russian, like other Slavonic languages, lacks a strict word order.

Rather than controlling for the variation introduced by relying on authentic data, which runs the risk of observing how a-typical language is processed, we preferred to embrace this variation and incorporate it as an integral element into the analysis by using regression modelling techniques, as explained in Sections 3.2 and 3.3.

Procedure

The experiments were run in a quiet room at the Institute for Linguistic Studies of the St Petersburg branch of the Russian Academy of Sciences. Participants provided personal information prior to attending using Google forms. They had also been sent information sheets and consent forms and were offered the opportunity to read those again at the testing location where the documents were also signed and handed over to the experimenter.
The self-paced reading task was programmed in PsychoPy. A Cedrus response pad was connected to a Windows 7 laptop (Intel i5 core) with Nvidia graphics card; all subjects completed the task individually on the same laptop. The presentation used a word-by-word template with no placeholders. The self-paced reading task was preceded by a serial reaction time task and followed a digit span task that are not described in this chapter. All subjects were debriefed after the session.

3.2 Data analysis: mixed effects linear regression model

A first set of models to explain the time reading the TRY verb was run using Generalized Linear Mixed effects Regression Modelling (GLMM) (e.g. Baayen et al. 2008), using R package lme4 (Bates et al. 2015). Generalized Linear Mixed effects regression Modelling is an extension of Generalized Linear Modelling so that the predictor effects are divided into fixed and random effects. Fixed effects represent the variables and their interactions we are interested in making inferences about, beyond our sample to the entire population. In contrast, random effects are variables which we presume to represent the effects of/in gathering the (random) sample we have (such as individual differences between speakers, experimental factors that may not be representative of the entire population of speakers or of the phenomenon of interest in its entirety), and the distorting impacts of which we want to minimize when drawing inferences about the fixed effects.

During the data preparation stage, some observations were excluded for the following reasons: data from 2 female subjects had to be discarded because the software crashed half-way through the reading task; 3 stimuli were excluded because they used a periphrastic future, which removed the tense marking from the TRY verb. Following standard procedure all responses were excluded that took less than .05 seconds and were more than 2 standard deviations removed from the mean. Baayen & Milin (2010) discuss some of the implications of this approach, and we mention explicitly here that our findings change with a more cautious trimming of datapoints as advocated in Baayen & Milin (2010). Numerical variables were log-transformed.

Probability of occurrence as specified by the regression model described in Section 2 was predicted by either the (logarithm of the) reading time on the TRY verb, the (logarithm of the) reading time on the word following the TRY verb (i.e. the infinitive) or the residualized (logarithm of the) reading time on the infinitive; the latter cases can be considered as “spill-over” effects of reading the TRY verb. Control variables were length of the word, position of the word (in the sentence) and (except in the residualized model) reading time on the previous word. Two variants were run of each model: one in which control variables were included in the fixed effects structure, and one whereby they were included in the random

effects structure. The latter approach gives us an idea of what factors in general seem to affect reading time, whereas the former tries to focus on the effect of prescribed factors in the apparent immediate textual, linguistic context, with the effect of the random factors "neutralized"; below we show the results of the latter models. The random effects structure always included at least subject, item and index.

The base model with probability of occurrence predicting reading time on the TRY verb showed a negative correlation between reading time and probability with higher probability TAM combinations facilitating reading. In other words, TRY verbs were read slightly more quickly if encountered in their expected form. Yet, as soon as variables were entered that are standardly used as control variables in the analysis of reading times data, in particular length of the TRY verb and position of the TRY verb in the sentence, the inverse correlation effect between reading time and probability, though still seemingly present, became overshadowed by the effects of the control variables and therefore non-significant.

Linear mixed model fit by REML
Formula: log(RT) ~ Probability + (1 | Participant) + (1 | TRY verb) + (1 | Position of TRY verb in sentence) + (1 | Length of TRY verb) + (1 | Position of sentence in experiment)

Random effects:

Groups	Name	Variance	Std.Dev.
Participant	(Intercept)	1.0302e-01	0.32097289
Sentence position	(Intercept)	2.6211e-03	0.05119630
Verb position	(Intercept)	2.7466e-03	0.05240846
Verb Length	(Intercept)	3.6840e-04	0.01919366
TRY verb	(Intercept)	3.0275e-07	0.00055022
Residual		4.1120e-02	0.20277984

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	-0.31577	0.05756	-5.486
Probability	-0.01277	0.03232	-0.395

There was some evidence of a spill-over effect as there was a stronger negative correlation with reading time on the word following the TRY-verb, which is typically the infinitive. Although a stronger effect on the infinitive would be expected on linguistic grounds (cf. Divjak 2004/2010 who showed that the distribution of TRY verbs is distinct from that of typical main verbs and instead resembles “light” verbs such as modals and phasals), this effect too failed to reach significance.

Linear mixed model fit by REML
Formula: log(RTspillOver) ~ Probability + (1 | Participant) + (1 | TRY verb) + (1 | Position of TRY verb in sentence) + (1 | Length of TRY verb) + (1 | Position of sentence in experiment)

Random effects:

Groups	Name	Variance	Std.Dev.
Participant	(Intercept)	0.11943454	0.345593
Sentence position	(Intercept)	0.00331632	0.057588
Verb position	(Intercept)	0.00068924	0.026253
Verb Length	(Intercept)	0.00000000	0.000000
TRY verb	(Intercept)	0.00000000	0.000000
Residual		0.04449880	0.210947

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	-0.31709	0.06000	-5.285
Probability	-0.02039	0.03169	-0.643

Overall, control variables such as the reading time of the previous word, the length of the TRY verb and the position of the sentence in the experiment all explain more about the speed with which the TRY verbs are read than does the probability of verb occurrence given the TAM marking.

In a third set of linear models residualization was performed because of autocorrelation issues: the (logarithm) of the reading time on the word preceding the TRY verb and the position of the TRY verb in the sentence account for 72% of the variance in the time spent reading the TRY verb. For the residualization, we used the (logarithm) of the (first pass) reading time on the word preceding the TRY verb and position of the verb in the sentence as predictors. Note that this dataset contained slightly fewer observations, 745 instead of 825, because in some cases the TRY verb was the first word in the sentence, meaning there is no preceding word with reading time, or because the reading time for the previous word was missing due to various reasons, e.g. having been skipped on first pass.

Linear mixed model fit by REML
Formula: logRTresid ~ Probability + (1 | Participant) + (1 | TRY verb) + (1 | Position of TRY verb in sentence) + (1 | Length of TRY verb) + (1 | Position of sentence in experiment)

Random effects:

Groups	Name	Variance	Std.Dev.
Participant	(Intercept)	0.00850915	0.092245
Sentence position	(Intercept)	0.00010236	0.010117
Verb position	(Intercept)	0.00086404	0.029395
Verb Length	(Intercept)	0.00193519	0.043991
TRY verb	(Intercept)	0.00000000	0.000000
Residual		0.05340367	0.231092

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	-0.0004919	0.0293385	-0.017
Probability	-0.0299761	0.0386436	-0.776

There are a number of possible explanations for the lack of significance where one would expect such on the basis of previous research. These explanations concern the phenomenon and the stimulus selection, the corpus data annotation, the experimental paradigm, the sample size and the assumptions underlying linear regression. We will address each of these points in turn.

First of all, different from other experimental work, the task our subjects faced was virtually "impossible": of the nine TRY verbs at our disposal, the three verbs we were targeting all belong to the same cluster (cf. Divjak 2003; Divjak & Gries 2006; 2008), meaning they are the most similar from among a group of 9 synonymous verbs. This makes it very hard to find properties that distinguish between these verbs. The already subtle differences were further obscured because we decided to work with authentic experimental items: all stimuli were attested in our corpus, meaning that all contexts we provided "fit" the verbs: some fit better, some fit worse (in comparison to the other two TRY verbs in that same context, or in comparison to all other contexts for that same TRY verb) but all are possible and had been genuinely produced by a speaker/writer.

Secondly, the corpus-model that produced the predicted probabilities did not know (or care) about verb length and position of the verb in the sentence; including this information might have changed the calculated probabilities. However, it could be argued that in the corpus sources which we used the authors had ample time to consider the composition of the sentences they wrote, thus being able to take into consideration the "linguistic goodness" of the entire context, i.e. the full sentence they were writing, including that part of the context following the TRY verb.

Thirdly, a word-by-word self-paced reading paradigm may be too mechanistic to pick up the subtle differences in reading times we are expecting; it is possible for subjects to fall into a pattern whereby the button presses guide reading speed rather than measure reading speed. It could therefore be suggested that eye-tracking would be a more suited experimental paradigm. Self-paced reading is, however, considered to be a robust technique, and reading latencies from both tasks correlate.[3]

Fourthly, given the subtlety of the effect, our sample size may have been too small: simulations show that we would need 100 times more data for the effect of probability on reading time to reach significance. This ties in with Jones and Tukey's (2000) criticism of null-hypothesis testing: given that eventually, we are likely to observe an effect, the question really becomes whether or not the effect is apparent enough in the data we have, and whether that effect is meaningful in terms of theory-based predictions and common sense.

A final explanation for the lack of a significance effect where one would expect such on the basis of previous research relates to the assumptions underlying the statistical modelling technique we used. The assumption underlying a GLMM is that the effects are linear and continuous; a linear model can fail to pick up significant effects that are non-linear in nature. In the next Section, we explore how such non-linearities in the data can be dealt with and modelled in more detail.

[3] Miwa et al (2014) report that in case of multiple fixations, typically, the later fixations tend to be more similar to lexical decision latencies.

3.3 GAMM Mixed effects additive models

In a second modelling round we considered a non-linear treatment of reading times using Generalized Additive Mixed Models (GAMM). GAMMs are an extension of the linear mixed model that make it possible to model a response variable as a nonlinear function of one or more predictor variables, using, e.g., thin plate regression splines. GAMMs have recently been applied successfully to linguistic and psycholinguistic data ranging from dialectometry (Wieling et al. 2011) to electromagnetic articulography (Tomaschek et al. 2014) and from EEG data (Kryuchkova et al. 2014) to pitch contours (Koesling et al. 2012).

The software available in the mgcv package for R by Wood (Wood 2006, Wood 2011) offers a wide range of statistical tools for the modelling of both fixed-effect factors, random-effects, covariates, and their interactions. Whereas the linear mixed model allows for the specification of a model in which a regression line Y = a + bX is modulated by Gaussian uncertainty for intercept and slope for a grouping factor F, effectively calibrating the regression line for each level of F, GAMMs offer the possibility to include a main effect of Y as a potentially nonlinear function of X, together with 'random nonlinear curves' for each level of F that are shrunk towards zero, under the constraint that these random curves have the same smoothing parameter. This is especially useful for modelling subject-specific variation in how participants perform in the course of an experiment. We ran GAMMS using package mgcv 1.8-5 (Wood 2015) in R 3.1.3 (March 2015).

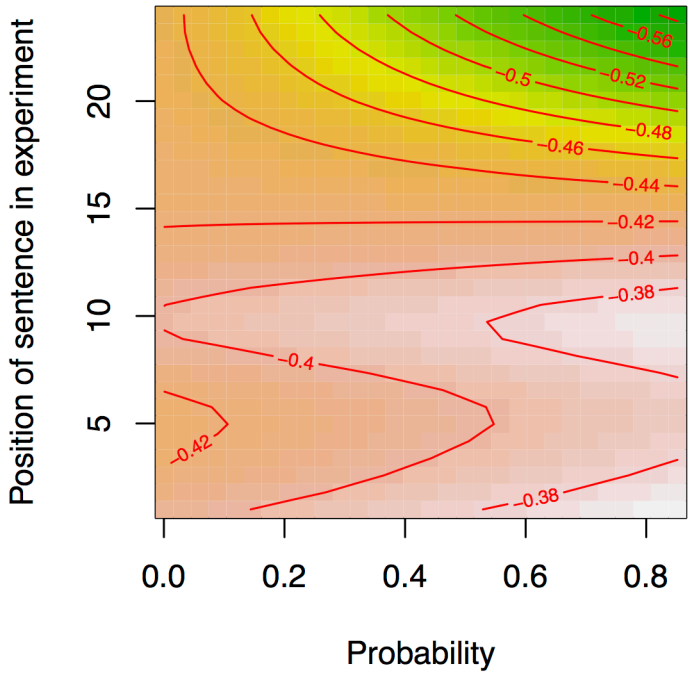
The specification for the best model, using the same dataset as for the GLMM models, is given below. It includes the length of the TRY verb as parametric coefficient, a smooth for the position of the sentence in the experiment, a factorial smooth for participant by position of the sentence in the experiment as well as an interaction between the probability of the verb given the TAM marking and the rank of the sentence in the experiment as tensor product. All terms contribute significantly to an explanation of the time it takes subjects to read the TRY verb in question.

```
Formula:
log(RT) ~ s(Position) + te(Probability, trial_order) + s(trial_order,
    participant, bs = "fs", m = 1) + CriticalLength

Parametric coefficients:
                 Estimate    Std. Error   t value     Pr(>|t|)
(Intercept)     -0.411802    0.068319     -6.028      2.64e-09 ***
CriticalLength   0.011211    0.004865     2.304       0.0215 *

Approximate significance of smooth terms:
                                edf     Ref.df      F       p-value
s(Position)                     4.420   5.296       5.517   3.96e-05 ***
te(Probability,trial_order)     4.224   4.649       7.007   5.32e-06 ***
s(trial_order,participant)      84.040  332.000     6.651   < 2e-16 ***

R-sq.(adj) =  0.737   Deviance explained = 76.7%
fREML = -54.623  Scale est. = 0.037683  n = 825
```

The significance of the interaction between probability and position of the sentence in the experiment confirms that TAM marking is picked up by native speakers and plays a role in on-line processes as captured by a self-paced reading task: words are read more quickly if presented with their distinctive TAM marking. Yet, the three-way interaction also signals that the readers' reaction to probability is not uniform throughout the task. In order to understand what is happening, we plot the interaction.

In Figure (1) the tensor interaction is plotted with probability on the X axis and the rank of the sentence in the experiment on the Y axis. The colours on the graph represent reading times and shade from pale pink in the bottom right corner over deep pink to yellow and green in the top right corner. The values on the isolines are the logarithms of the reading times. The value on the isoline around the white area in the bottom right corner being -0.38 (the logarithm of a reading time of 0.68 seconds) and the value on the isoline around the green area in the top right corner being -0.56 (the logarithm of a reading time of 0.57 seconds), white signals longer reading latencies and green shorter reading latencies.

Figure (1) Interaction of probability of TRY verb given TAM marking and rank of the sentence in the experiment in the GAM Model



The subjects' behaviour changes half-way through the experiment: while they start out reading slowly and in fact reading verbs with expected TAM marking slightly slower than verbs with unexpected TAM marking, they end the experiment reading quickly and reading verbs with expected TAM marking most quickly, as predicted. Although it is typical for subjects to change pace during an experiment (some become faster, others slower) the pattern exhibited by their reading behaviour does not display the U-shape often seen.

We have pointed out that there are two, competing, interpretations of reaction times. How do these account for the effect we observe? On the standard interpretation (Kaiser 2013) longer reading latencies signal processing difficulty, while shorter reading latencies signal processing ease. This would mean that our subjects found verbs with highly likely TAM

marking at first slightly more difficult to process than verbs with unlikely TAM markings, which goes against the majority of findings in this area. Given that there were three different sets of experimental items and that the order of the sentences was randomized for each subject, this effect cannot have been caused by inaccurate predictions of the regression model or unfortunate lexical side-effects in the stimuli used at the beginning of the experiment. Then why did our subjects slow down when encountering TRY verbs in their preferred form? The alternative interpretation of what reading times show suggests that a slowdown in reading speed signals a large change in surprisal (Hale 2003, Levy 2008). This would mean that our subjects were able to make sense of the sentence they were reading only when they reached a familiar item, i.e. the TRY verb in its preferred form. This seems like a more plausible interpretation: subjects were reading authentic, literary sentences, but without wider context. They did not know anything about the background of the situation described in the sentence they were reading and may well have clicked through the first words in the sentence, collecting information, until a familiar word appeared and they were in a position to start integrating that information. Since the subjects knew that some sentences were followed by a yes/no question that they needed to answer, it is plausible to assume that they will have paused to integrate information and to prepare for the question. The novelty of reading literary sentences without further context and the preparation in anticipation of what turned out to be relatively straightforward questions will have worn off over the course of the experiment and about half-way through (to be precise: after 5 training sentences, 12 filler items and 12 experimental items), subjects would have learnt to expect some kind of light verb expressing attempt or perception. This would have made it possible for them to start reading in a more natural way, skipping more quickly over expected words. This delay in effect may have been exacerbated by the fact that our subjects came from a prescriptive linguistic tradition and from an instruction-based rather than inquiry-based educational system. Both factors contribute to a desire to do well in test situations.

4. Looking back, looking forward and looking outward

The impetus for the research question discussed in this chapter comes from the morphological richness, typical of Slavic languages. The fact that the bulk of corpus- and psycholinguistic research is done on English, which is a morphologically poor language, has lead researchers to assume that the lemma level suffices as guide for annotation and for claims about representation. Data from a morphologically rich language like Russian with abundant inflectional markings show that this finding is a side-effect of the properties of the language studied: if inflectional markings are present, they are detected and used by speakers in on-line processing, as witnessed by an increase in reading speed on TRY verbs carrying their distinctive TAM marking. This finding highlights the extent of the knowledge speakers have of distributional patterns typical of their language and encourages linguists to think of language less in terms of inventories of items that can be freely combined in an unlimited number of ways, but in terms of prefabricated chunks that are more or less expected given the context.

The recognition of the existence of degrees of meaning difference and the interest in pinpointing the source of these fine-grained differences is clearly reminiscent of the Russian semantic tradition. This interest underlies the corpus analysis on which the on-line

experiment was built. Advantages of basing experiments on extensive corpus-linguistic research include having access to a very rich knowledge base that increases one's chances of avoiding confounds due to knowing what to expect and having good theoretically motivated and empirically supported idea why that should be so. Admittedly, a corpus-based approach of the type described in this chapter is a labour intensive endeavour that furthermore complicates the experimental design and makes the use of advanced statistical methods necessary. Yet, at the same time, it is economical in that it captures subjects' reactions to a much wider range of possible contexts (i.e. variable combinations) than is typically the case in an experiment, and it provides a very high level of control, not only over the frequency with which the words in the sentences occur, but also – and crucially – over the likelihood of having words co-occur with each other. That being said, one piece of information that corpus linguists should consider including in their models is the position of the word of interest in the sentence: information may well be structured differently for different verbs.

Considering the relation between a verb and the way in which information is structured is especially important if the preceding context will be cut out in the experimental presentation, such as in a linear reading experiment with piecemeal exposure. The word-by-word presentation is not natural and the clicking might pace reading speed rather than record it. Nevertheless, the task is relatively intuitive for subjects and avoids uncertainty about what is causing longer reading time when chunks are presented (looking back/forward). Although self-paced reading is a robust task yielding reaction times, which are a type of data about which much is known and which are strongly correlated with (first pass) reading times from eye-tracking, the ecological validity of on-line tasks would benefit from including information structure in experimental stimuli and presenting the words of interest in the position in the sentence which seems most natural for them and in a larger context: even if the sentences themselves are authentic and extracted from a corpus, having a TRY verb without preceding context may well be unnatural as there is no information on who is trying, why, and what (cf. Roland & Jurafsky 2002)? On a practical level, our data shows that subjects' prior experience and cultural background need to be taken into account when running experiments: subjects who are not accustomed to (psycholinguistic) experiments, who come from prescriptive linguistic traditions and/or from instruction- rather than inquiry-based educational systems may need to be given a longer time to practice to overcome subconscious barriers and start to show natural behaviour. Unless we address these issues, we risk continuing to measure what people do when things are not as they normally are and we might miss an effect that is indeed present in the data.

In our case, we needed a powerful statistical technique, generalized additive mixed effects regression modelling, to detect the expected relation between probability and reading time. Yet, the algorithms underlying standard statistical classifiers such as regression techniques were not designed to mimic human learning. Although they show good prediction accuracy, the drawback is that they yield cognitively unrealistic models that are of limited interest to usage-based linguistics from a theoretical point of view. Are probabilities the proper constructs to capture the processes that are at work? Research in progress (Divjak et al.) re-models this same data using a biologically and cognitively plausible model of learning, the Naïve Discriminative Learner (Baayen et al. 2011) to answer a number of questions raised in this chapter. First, is it truly the probability of an abstract semantic category that is driving the behaviour that we see in this self-paced reading task, or is it the distinctive cues in the orthographic input that support TAM? Second, if it is a probability, how would that probability be learned? Could an approach based on discrimination learning shed light on

this process? Third, why is there an interaction between probability and position of the sentence in the experiment? Doesn't this suggest learning in the course of the experiment? And if so, what are our subjects learning? And finally, how can we obtain further insight into and evidence for the Hale (2003)/Levy (2008) interpretation of a slow-down in reading latencies that our data seems to be supporting, but that has not been the dominant interpretation in the literature? We leave these questions to future research.

References

Ambridge, Ben, Julian M. Pine, Caroline F. Rowland, and Franklin Chang. 2012. The roles of verb semantics, entrenchment, and morphophonology in the retreat from dative argument-structure overgeneralization errors. *Language* 88 (1): 45-81.

Antić, Eugenia. 2012. Relative frequency effects in Russian morphology. In Stefan Th. Gries & Dagmar Divjak, eds. *Frequency effects in language learning and processing*. *Vol. 1.* Berlin, Boston: De Gruyter Mouton, 83-108.

Apresjan, Jurij D. 1995 [1974]. *Избранные труды. Том I. Лексическая семантика: синонимические средства языка.* [Selected works. Volume I. Lexical semantics: The synonymic means of language]. Moskva: Skola “Jazyki Russkoj Kul’tury”.

Apresjan, Jurij D et al. 1999. *Новый объяснительный словарь синонимов русского языка*. *Vol. 1.* Moskva: Škola “Jazyki Russkoj Kul’tury”.

Arppe, Antti. 2008. Univariate, Bivariate and Multivariate Methods in Corpus-Based Lexicography – A Study of Synonymy. PhD Dissertation. Publications of the Department of General Linguistics, University of Helsinki. [https://helda.helsinki.fi/handle/10138/19274, last accessed on 28/05/2015]

Antti Arppe. 2013a. Package polytomous: Polytomous logistic regression for fixed and mixed effects. R package version 0.1.6. http://CRAN.R-project.org/package=polytomous

Arppe, Antti. 2013b. Extracting exemplars and prototypes. R vignette to accompany Divjak & Arppe. 2013. Available from [http://cran.r-project.org/web/packages/polytomous/vignettes/exemplars2prototyp es.pdf]

Arppe, Antti and Juhani Järvikivi. 2007. Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3 (2): 131-159.

Baayen, R. Harald, Doug J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59 (4): 390–412 (Special issue on Emerging Data Analysis Techniques).

Baayen, R. Harald, Ton Dijkstra, and Robert Schreuder. 1997. Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language* 36: 94-117.

Baayen, R. Harald and Petar Milin. 2010. Analyzing reaction times. International Journal of Psychological Research, 3.2, 12-28.

Baayen, R. Harald, Petar Milin, Dušica F. Đurđević, Peter Hendrix, and Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review*, 118(3): 438-481

Baayen, R. Harald, Laura A. Janda, Tore Nesset, Stephen Dickey, Anna Endresen, and Anastasija Makarova. 2013. Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics* 37, 253-291

Bates, Douglas, Martin Maechler, Ben Bolker, and Steve Walker 2015. *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-8, <URL: http://CRAN.R-project.org/package=lme4>.

Backhaus, Klaus, Bernd Erichson, Wulff Plinke, and R. Weiber. 1996. *Multivariate Analysemethoden: eine anwendungsorientierte Einführung. 8. ed.* Berlin/Heidelberg/NewYork: Springer.

Biber, Douglas, Susan Conrad and Randi Reppen. 1998. *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld, eds. *Roots: Linguistics in Search of Its Evidential Base*. Berlin: Mouton de Gruyter, 77–96.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the Dative Alternation. In Gerlof Bouma, Irene Krämer & Joost Zwarts, eds. *Cognitive Foundations of Interpretation*. Amsterdam: Royal Netherlands Academy of Science, 69–94.

Bresnan Joan and Marilyn Ford. 2010. "Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English." *Language* 86(1): 186--213.

Church, Kenneth Ward, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In U. Zernik (ed.). *Lexical acquisition: Exploiting on-line resources to build a lexicon*, 115–164. Hillsdale, NJ: Lawrence Erlbaum.

Church, Kenneth Ward, William Gale, Patrick Hanks, Donald Hindle, and Rosamund Moon. 1994. Lexical substitutability. In B.T.S. Atkins and A. Zampolli (eds.). *Computational approaches to the lexicon*, 153–177. Oxford and New York: Oxford University Press.

Cruse, D. Alan. 2000. *Meaning in Language: an Introduction to Semantics and Pragmatics.* Oxford: Oxford University Press.

Derman, Emanuel. 2011. Models. Behaving. Badly.: Why Confusing Illusion with Reality Can Lead to Disaster, on Wall Street and in Life. New York: Free Press

Divjak, Dagmar. 2003. On trying in Russian: a tentative network model for near(er) synonyms. In "Belgian Contributions to the 13th International Congress of Slavicists, Ljubljana, 15-21 August 2003". Special Issue of Slavica Gandensia 30: 25-58.

Divjak, Dagmar. 2004. Degrees of Verb Integration.Conceptualizing and Categorizing Events in Russian. Ph.D. diss., Dept. of Oriental & Slavic Studies. K.U.Leuven (Belgium).

Divjak, Dagmar. 2006 Ways of Intending: Delineating and Structuring Near-Synonyms. In: Gries, St. T. & Stefanowitsch, A. (eds.) Corpora in cognitive linguistics. Corpus-based Approaches to Syntax and Lexis. Berlin-New York: Mouton de Gruyter, 19-56. [Trends in Linguistics 172].

Divjak, Dagmar. 2010. Structuring the lexicon: a clustered model for near-synonymy. Berlin, New York: Mouton de Gruyter. [Cognitive Linguistics Research]

Divjak, Dagmar and Antti Arppe. 2013. Extracting prototypes from exemplars. What can corpus data tell us about concept representation? *Cognitive Linguistics* 24 (2): 221–274.

Divjak, Dagmar, Ewa Dąbrowska, and Antti Arppe. 2016. Machine meets man: Evaluating the psychological reality of corpus-based probabilistic models. *Cognitive Linguistics* 27(1).

Divjak, Dagmar, and Stefan Th. Gries. 2006. Ways of Trying in Russian. Clustering Behavioral

Profiles. In: Journal of Corpus Linguistics and Linguistic Theory 2 (1): 23-60.
Divjak, Dagmar, and Stefan Th. Gries. 2008. Clusters in the Mind? Converging evidence from near-synonymy in Russian. *The Mental Lexicon* 3 (2): 188-213.
Divjak, Dagmar, Petar Milin, and R. Harald Baayen. (in progress). Reading Russian synonyms. (working title).
Edmonds, Philip, and Graeme Hirst. 2002. Near-synonymy and Lexical Choice. Computational Linguistics, 28 (2): 105-144.
Erker Daniel and Gregory R. Guy. 2012. The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. *Language* 88 (3): 526-557.
Field, Andy, Jeremy Miles, and Zoe Field. 2012. Discovering Statistics Using R. London: Sage publications.
Firth, J. Rupert. 1957. *Papers in Linguistics 1934–1951.* London: Oxford University Press.
Ford, Marilyn and Joan Bresnan. 2013a. `They whispered me the answer' in Australia and the US: A comparative experimental study. In Tracy Holloway King & Valeria de Paiva, eds. *From Quirky Case to Representing Space: Papers in Honor of Annie Zaenen.* Stanford: CSLI Publications, 95-107. (http://web.stanford.edu/group/cslipublications/cslipublications/Online/azfest-final.pdf, last accessed on 22/01/2015).
Ford, Marilyn & Joan Bresnan. 2013b. Using convergent evidence from psycholinguistics and usage. In Manfred Krug & Julia Schlüter, eds. *Research Methods in Language Variation and Change*. Cambridge: Cambridge University Press, 295-312.
Geeraerts, Dirk. 1985. Preponderantieverschillen bij bijna-synoniemen. In: *De nieuwe taalgids* 78, 18-27.
Gries, Stefan Th. 2003. *Multifactorial analysis in corpus linguistics: a study of Particle Placement*. London & New York: Continuum Press.
Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: the many meanings of *to run*. In Stefan Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*, 57-99. Berlin & New York: Mouton de Gruyter.
Gries, Stefan Th. & Martin Hilpert. 2010. Modeling diachronic change in the third person singular: a multifactorial, verb-and author-specific exploratory approach. *English Language and Linguistics* 14 (3): 293-320.
Grondelaers, Stefan & Dirk Speelman. 2007. A variationist account of constituent ordering in presentative sentences in Belgian Dutch. *Corpus Linguistics and Linguistic Theory* 3 (2): 161-193.
Hale J. 2003. The information conveyed by words in sentences. *Journal of Psychological Research* 32: 101–123.
Hanks, Patrick 1996. Contextual Dependency and Lexical Sets. International Journal of Corpus Linguistics, Vol. 1 , No. 1, pp. 75-98.
Harris, Zellig. 1954. Distributional structure. *Word* 10(23). 146-162.
Inkpen, Diana and Graeme Hirst 2006. Building and Using a Lexical Knowledge- Base of Near-Synonym Differences. Computational Linguistics 32:2 (June 2006), pp. 223-262.
Jones, Lyle V. & John W. Tukey. 2000. A sensible formulation of the significance test. Psychological Methods. 2000 Dec 5(4): 411-4.
Kaiser, E. (2013). Experimental paradigms in psycholinguistics. In: R. J. Podesva & D. Sharma (eds.), *Research Methods in Linguistics* (pp. 135-168). Cambridge: Cambridge University Press.

Kendall, Tyler, Joan Bresnan & Gerard Van Herk. 2011. The dative alternation in African American English: Researching syntactic variation and change across sociolinguistic datasets. *Corpus Linguistics and Linguistic Theory* 7 (2): 229-244.

Kjellmer, Göran. 2003. Synonymy and corpus work: on almost and nearly**.** *ICAME Journal* 27: 19-27.

Klavan, Jane 2012. Evidence in linguistics: corpus-linguistic and experimental methods for studying grammatical synonymy (Dissertationes Linguisticae Universitatis Tartuensis). Tartu: University of Tartu Press.

Klavan, J. & D. Divjak. (in press for 2016). Review article: The Cognitive Plausibility of Statistical Classification Models: Comparing Textual and Behavioral Evidence. Special issue of *Folia Linguistica*, edited by Martin Hilpert, Karolina Krawczak and Małgorzata Fabiszak.

Koesling, K., Kunter, G., Baayen, R., & Plag, I. (2012). Prominence in triconstituent compounds: Pitch contours and linguistic theory. Language and Speech, 56 (4), 529-554.

Kostić, Aleksandar, and Jelena Havelka (2002). Processing of verb tense. *Psihologija* 35.3-4:299-316.

Kostić, Aleksandar, and Jelena Mirković (2002). Processing of inflected nouns and levels of cognitive sensitivity. *Psihologija* 35.3-4:287-297.

Kryuchkova, T., Tucker, B. V., Wurm, L., & Baayen, R. H. (2012). Danger and usefulness in auditory lexical processing: evidence from electroencephalography. Brain and Language, 122 , 81–91.

Levy, R. 2008. 2008. Expectation-based syntactic comprehension. *Cognition* 106: 1126–1177.

Miwa, K., Garry Libben, Ton Dijkstra, and R. Harald Baayen. 2014. The time-course of lexical activation in Japanese morphographic word recognition: Evidence for a character-driven processing model. *The Quarterly Journal of Experimental Psychology*, 67: 79-113.

Mondry, Henrietta and John Taylor. 1992. On lying in Russian. In: *Language & Communication* 12 (2): 133-143.

Newman, John. 2008. Aiming low in linguistics: Low-level generalizations in corpus-based research [downloaded from http://www.johnnewm.org/downloads/; page last accessed on 07.06.2013]

Perek, Florent. 2015. *Argument structure in usage-based Construction Grammar: Experimental and corpus-based perspectives.* Amsterdam: John Benjamins.

Raymond, William D. & Esther L. Brown. 2012. Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish. In Stefan Th. Gries & Dagmar Divjak, eds. *Frequency effects in language learning and processing*. Berlin: Mouton de Gruyter, 35-52.

Roland, Douglas and Daniel Jurafsky. 2002. Verb Sense and Verb Subcategorization Probabilities. In Stevenson, Suzanne and Paola Merlo (eds.), *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*. Amsterdam/Philadelphia: John Benjamins, 325-346.

Schmid, Hans-Jörg. 1993. *Cottage und Co., idea, start vs. begin: die Kategorisierung als Grundprinzip einer differenzierten Bedeutungsbeschreibung.* Tübingen: Niemeyer.

Sinclair, John. 2001. *Corpus, Concordance, Collocation.* Oxford: Oxford University Presss.

Solovyev,Valerij D. and V.P. Bajraševa. 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.

Szmrecsanyi, Benedikt. 2013. Diachronic Probabilistic Grammar. *English Language and Linguistics* 1 (3): 41-68.

Taylor, John. 2003. Near synonyms as co-extensive categories: 'high' and 'tall' revisited. In: *Language Sciences,* 25: 263-284.

Theil, Henri. 1970. On the Estimation of Relationships Involving Qualitative Variables. *The American Journal of Sociology* 6 (1): 103-154.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work.* [Studies in Corpus Linguistics 6]. John Benjamins: Amsterdam/Philadelphia.

Tomaschek, Fabian, Ben Tucker, Martijn Wieling, and Harald Baayen. 2014. Vowel articulation affected by word frequency. In *Proceedings of 10th ISSP*, Cologne, 429–432.

Wieling, Martijn, John Nerbonne, and Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, 6 (9), e23613.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach, and Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30 (3): 382-419.

Wood, Simon. 2006. *Generalized Additive Models. An Introduction with R.* CRC Press: Boca Raton.

Wood, Simon. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B), 73: 3–36.

Wood, Simon. 2015. Package mgcv. R package version 1.8-5. Available from https://cran.r-project.org/web/packages/mgcv/index.html.