

## ***Cognitive Linguistics* 2016 1 (27)**

### **Machine Meets Man: evaluating the psychological reality of corpus-based probabilistic models**

Dagmar Divjak, The University of Sheffield  
Ewa Dąbrowska, Northumbria University  
Antti Arppe, University of Alberta

#### **Abstract**

Linguistic convention allows speakers various options. Evidence is accumulating that the various options are preferred in different contexts yet the criteria governing the selection of the appropriate form are often far from obvious. Most researchers who attempt to discover the factors determining a preference rely on the linguistic analysis and statistical modeling of data extracted from large corpora.

In this paper, we address the question of how to evaluate such models and explicitly compare the performance of a statistical model derived from a corpus with that of native speakers in selecting one of six Russian TRY verbs. Building on earlier work by Divjak (2003, 2004, 2010) and Divjak & Arppe (2013), we trained a polytomous logistic regression model to predict verb choice given the context. We compare the predictions the model makes for 60 unseen sentences to the choices adult native speakers make in those same sentences.<sup>1</sup> We then look in more detail at the interplay of the contextual properties and model computationally how individual differences in assessing the importance of contextual properties may impact the linguistic knowledge of native speakers. Finally, we compare the probability the model assigns to encountering each of the 6 verbs in the 60 test sentences to the acceptability ratings the adult native speakers give to those sentences. We discuss the implications of our findings for both usage-based theory and empirical linguistic methodology.

#### **Acknowledgments**

The experiment received ethical approval from the University of Sheffield, School of Languages & Cultures; the data were collected in 2013. The financial support of the University of Sheffield in the form of a 2013 SURE summer research internship to Clare Gallagher is gratefully acknowledged; Clare set up the acceptability ratings study and recruited participants for this task.

---

<sup>1</sup> Note that we use the word *predict* in the statistical sense, i.e., “identify as the most likely choice, given the data the model was trained on”.

## Introduction

A particular idea can often be coded linguistically in several different ways: that is to say, linguistic convention allows speakers various options. At the lexical level, speakers can choose from sets of near synonyms (*walk, march, stride, strut...*). Similarly, at the grammatical level, there are often several options for encoding slightly different construals of the same situation: for instance, in English, there are several ways of marking past events (*was walking, walked, had walked*), two indirect object constructions (*give him the book vs give the book to him*), and so on. Cognitive linguists have long been claiming that languages abhor (complete) synonymy and evidence is accumulating showing that in the vast majority of cases, the various options are preferred in different contexts.

However, the criteria governing the selection of the appropriate form are often far from obvious, and hence, there is now a considerable amount of empirical work attempting to describe the differences between near synonymous lexemes or constructions (for book-length treatments see Arppe 2008, Divjak 2010, Klavan 2012 and references therein). Most researchers who attempt to discover the factors influencing a speaker's decision to use a particular form rely on the analysis of large corpora. A typical analysis involves extracting a large number of examples from a corpus and coding them for a number of potentially relevant features (Klavan 2012) or even as many potentially relevant features as possible (Arppe 2008, Divjak 2010). The usage patterns obtained can then be analyzed statistically to determine which of the candidate features are predictive of the form which is the focus of the study. The most rigorous studies also fit a statistical model to the data and test it on a new set of corpus examples (the testing set) to see how well it generalizes to new data.

One problem faced by researchers in this area is how to evaluate such models. A model that supplies the target form 85% of the time may be regarded as better than one that predicts it 80% of the time – but can this be regarded as adequate? After all, such a model still gets it wrong 15% of the time! The answer, of course, depends partly on (1) how many options there are to choose from (51% correct is very poor if there are only two options, but would be impressive if there were ten), but also on (2) the degree to which the phenomenon is predictable (100% correct is not a realistic target if the phenomenon is not fully predictable), as well as (3) what is being predicted: individual choices or rather proportions of choices over time. As Kilgariff (2005) and many others have observed: language is never ever random; however, it is also rarely, if ever, fully predictable.

The obvious solution for cognitive linguists is to compare the model's performance to that of native speakers of the language. Such a comparison could, in principle, result in three possible outcomes. First, the model may perform less well than humans. If this is the case, then the model is clearly missing something, and this tells us that we must go back to the data and find out what we have not coded for, add new predictors to the model, and test it again. Secondly, the model may perform as well as humans. This is clearly an encouraging outcome, but if we are interested in developing a psychologically realistic model (as opposed to simply describing the corpus data), we would want to make sure that the model is relying on the same criteria as the speakers. We could conclude that this was the case if the pattern of performance was similar, that is to say, if the model gives clear predictions (i.e., outputs a high probability for one particular option) when the speakers consistently choose the same option, and, conversely, if uncertainty in the

model (several options with roughly equal predicted probabilities, of e.g. 0.2-0.3 in the case of 3-5 alternatives, as opposed to one clear favourite) corresponded to variability in human responses. Finally, the model may perform better than humans. Statistical models have been found to outperform human experts in a number of areas including medical diagnosis, academic and job performance, probation success, and likelihood of criminal behaviour (Dawes, Faust and Meehl 1989, Grove et al. 2000, Stanovich 2010). To our knowledge, no model of linguistic phenomena currently performs better than humans (for instance, is able to choose the form that actually occurred in a particular context in a corpus more accurately than the average human informant) but it is perfectly possible that, as our methods improve, such models will be developed.

## 1. Previous studies

There are now a number of published multivariate models that use data, extracted from corpora and annotated for a multitude of morphological, syntactic, semantic and pragmatic parameters, to predict the choice for one morpheme, lexeme or construction over another. However, most of these studies are concerned with phenomena that involve binary choices (Gries 2003, De Sutter et al. 2008) and only a small number of these<sup>2</sup> corpus-based studies have been cross-validated (Keller 2000, Sorace & Keller 2005, Wasow & Arnold 2003, Roland et al. 2006, Arppe & Järvikivi 2007, Divjak & Gries 2008).<sup>3</sup> Of these cross-validated studies, few have directly

---

<sup>2</sup> There are a number of early studies that employ multiple explanatory variables but do not use these to construct multivariate models. Instead, they consider all possible unique variable-value combinations as distinct *conditions* (e.g. Gries 2002, Featherston 2005).

<sup>3</sup> Note that Grondelaers & Speelman (2007) and Kempen & Harbusch (2005) work the other way around and validate and refine experimental findings using corpus data.

evaluated the prediction accuracy of a complex, multivariate corpus-based model on humans using authentic corpus sentences (with the exception of Bresnan 2007, Bresnan & Ford 2010, Ford & Bresnan 2012, Ford & Bresnan 2013), and even fewer have attempted to evaluate the prediction accuracy of a polytomous corpus-based model in this way (but see Arppe & Abdulrahim 2013 for a first attempt). Below we will review the latter two types of cross-validated studies.

Bresnan (2007) was the first to evaluate a multivariate corpus-based model (Bresnan et al. 2007) designed to predict the binary dative alternation. A scalar rating task was used to evaluate the correlation between the naturalness of the alternative syntactic paraphrases and the corpus probabilities. Materials consisted of authentic passages attested in a corpus of transcriptions of spoken dialogue; the passages were randomly sampled from the centers of five equally sized probability bins, ranging from a very low to a very high probability of having a preposition dative construction. For each sampled observation the alternative paraphrase was constructed. Both options were presented as choices in the original dialogue context. Contexts were only edited for readability by shortening and by removing disfluencies. Items were pseudo-randomized and construction choices were alternated to make up a questionnaire. Each of the 19 subjects received the same questionnaire, with the same order of items and construction choices. Subjects were asked to rate the naturalness of alternatives in a given context by distributing 100 points over both options. Responses were analysed as a function of the original corpus model predictor variables by using mixed effects logistic regression. Bresnan found that subjects' scores of the naturalness<sup>4</sup> of the alternative syntactic paraphrases correlate well ( $R^2 = 0.61$ )

---

<sup>4</sup> Arppe & Järviö (2007) criticize Bresnan's set-up of operationalizing *naturalness* as a zero-sum game, with naturalness between the two alternatives always adding up to the same value, i.e. 100, as

with the corpus probabilities and can be explained as a function of the same predictors. Individual speakers' choices matched the choice attested in the corpus in 63% to 87% of all cases (with a baseline of 57% correct by always choosing the most frequently occurring option). Bresnan concluded that language users' implicit knowledge of the dative alternation in context reflects the usage probabilities of the construction.

Ford & Bresnan (2010, 2012, 2013) investigated the same question across American and Australian varieties of English. Relevant here is that they ran a continuous lexical decision task (Ford 1983) to check whether lexical-decision latencies during a reading task reflect the corpus probabilities. In a continuous lexical decision task subjects read a sentence word by word at their own pace, and make a lexical decision as they read each word (participants are presented with a sentence one word at a time and must press a "yes" or "no" button depending on whether the "word" is a real word or a non-word). The participants were instructed to read the contextual passage first and then make a lexical decision for all words from a specific starting point. That starting point was always the word before the dative verb. There were 24 experimental items, chosen from the 30 corpus items used in the scalar rating task (Bresnan 2007). A mixed effects model fit to the data confirmed that lexical-decision latencies during a reading task reflect the corpus probabilities: more probable sentence types require fewer resources during reading, so that RTs measured in the task decrease in high-probability examples.

---

their own study shows that even strong differences in terms of preference might nevertheless exhibit relatively small differences in *acceptability*. However, Bresnan's results would seem to indicate that the human participants were agreeing with the corpus-based estimates of the *proportions* of choice (in the long run) between the two alternatives (rather than with their *naturalness*). Of course, we cannot be sure what participants in an experiment are doing, regardless of how the instructions are formulated (cf. Penke & Rosenbach 2004).

Arppe & Abdulrahim (2013) contrast corpus data and force-choice data on 4 near-synonymous verbs meaning *come* in Modern Standard Arabic to assess the extent to which regularities extracted from a corpus overlap with collective intuitions of native speakers. A model of the corpus data was built using polytomous logistic regression based on the one-vs-all heuristic (Arppe 2008, 2013a) and was compared to data from a forced-choice task completed by 30 literate Bahraini native speakers of Arabic who read 50 sentences and chose the missing verb from a given list of verbs. The 50 experimental stimuli were chosen to represent the full breadth of contextual richness in the corpus data and the entire diversity of probability distributions, ranging from near-categorical preferences for one verb to approximately equal probability distributions for all four verbs. Arppe & Abdulrahim (2013) found that as the probability of a verb, given the context, rises, so does the proportion of selections of that verb in the context in question (*proportion* being the relative number of participants selecting the particular verb). Importantly there are hardly any cases where a low-probability verb would have received a high proportion of choices, and only a few in which high-probability verb would have received a low proportion of choices.

## **2. Russian verbs of trying**

In this paper, we explicitly compare the performance of a statistical model derived from a corpus with that of native speakers. The specific phenomenon that we will investigate concerns six Russian verbs (*probovat'*, *silit'sja*, *pytat'sja*, *norovit'*, *starat'sja*, *poryvat'sja*) which are similar in meaning – they can all be translated with the English verb *try* – but which are not fully synonymous. As explained in Divjak

(2010: 1-14), these verbs were selected as near-synonyms on the basis of a distributional analysis in the tradition of Harris (1954) and Firth (1957), with meaning construed as contextual in the Wittgensteinian sense. Synonymy was thus operationalized as mutual substitutability or interchangeability within a set of constructions, forming a shared constructional network. This is motivated by a Construction Grammar approach to language in which both constructions and lexemes are considered to 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.

To study verbal behavior within a shared constructional frame we build on earlier work by Divjak (2003, 2004, 2010), who constructed a database containing 1351 tokens of these verbs. Source of the data were the Amsterdam Corpus, supplemented with data from the Russian National Corpus, which contains written literary texts. About 250 extractions per verb were analysed in detail, except for *poryvat'sja*, which 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 verbs is so large (see Table 6 below) that manually annotating a sample in which the verbs would be represented proportionally would be prohibitively expensive. The sentences containing one of the six TRY verbs were manually annotated for a variety of morphological, semantic and syntactic properties, using the annotation scheme proposed in Divjak (2003, 2004). The tagging scheme 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. There were a total of 14 multiple-category variables amounting to 87 distinct variable categories or contextual properties.

Divjak and Arppe (2013) used this dataset to train a polytomous logistic regression model (Arppe 2013a, 2013b) predicting the choice of verb. 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 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 categories were retained (11 semantic and 7 structural), belonging to 7 different types. These are listed in Table 1.

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

Table 1. *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 probability for each verb in each sentence. More interestingly from an analyst's perspective, the model tells us how strongly each feature 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 preferences (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 was 51.7% (50.3% when tested on unseen data). This is well above chance: since there are six verbs, chance performance would have been at 16.7%. This overall accuracy may, however, still seem disappointingly low until we remember that the verbs have very similar meanings and are often interchangeable: that is to say, most contexts allow several, if not all, verbs. So the more interesting question is how the model's performance compares with that of humans. We explore this question in three studies.

### **3. STUDY 1 – FORCED CHOICE TASK**

In this study, we investigate Russian speakers' preferences for verbs of trying in specific sentential contexts using a force-choice task. We then compare the speakers' preferences to those of the model, assuming that the model "prefers" the verb with the highest predicted probability. Obviously choosing a verb to go in a

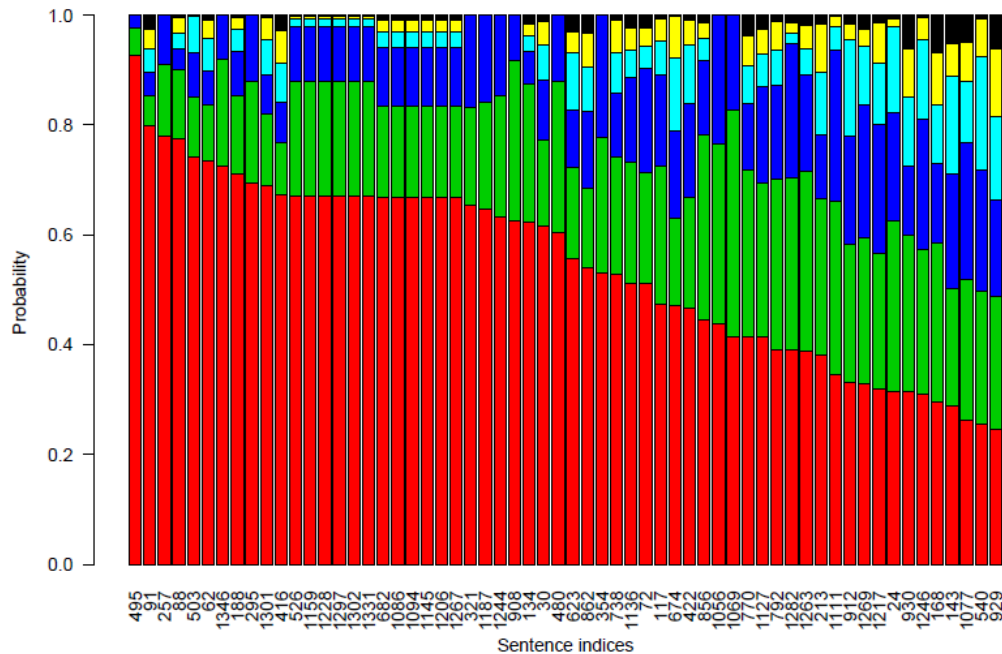
particular sentence is a fairly artificial task: it is not what speakers do during normal language use. However, a force-choice task provides useful information about speakers' preferences, and for this reason such tasks are routinely used in psycholinguistic research as well as in language testing. From our point of view, its major advantage is that it allows us to obtain comparable data from the model and from native speakers.

### **3.1. Method**

#### *3.1.1. Materials*

We extracted 60 sentences from the Divjak (2010) dataset. The sentences were selected to represent the whole spectrum of the probability distributions. The probabilities of the selected sentences are visualized in Figure (1) where each shade of grey represents a different verb (shades represent sentence-specific probabilities rather than verbs, i.e. dark grey is always used to mark the verb that has the highest probability of occurring, regardless of which of the six verbs it is; a lighter shade of grey is always used for the second most likely verb, etc.) and the height of each coloured portion of the bar represents the probability of the verbs occurrence as predicted by the Divjak and Arppe (2013) model. As we move from left to right, we see that the predominance of one verb over all other options diminishes, until we end at the right hand side with a number of cases in which the distribution of probabilities starts to equal out over all 6 verbs.

Figure 1. Probability distribution for TRY verbs across the 60 sentences



Four of the sentences were close to categorically biasing contexts according to the Divjak and Arppe (2013) model, i.e., the model assigned a probability of 0.70 or above to one verb, and the predicted probabilities for all other verbs were  $\leq 0.10$ . Thirty-one experimental sentences were strongly biasing, i.e., the model predicted a probability value of more than 0.50 for one of the verbs. In the remaining 25 sentences, there was no clear winner, with up to five verbs with predicted probabilities  $\geq 0.10$ . Because the sentence selection was driven by concerns about probability distribution, not all 6 target verbs are represented in equal numbers in the experimental sentences. Table 2 specifies the number of sentences used for each of the six TRY verbs.

<b>Verb</b>	<b>Sentences</b>
norovit'	12
probovat'	8
silit'sja	6
poryvat'sja	4
pytat'sja	10
starat'sja	20

Table 2. *Number of sentences per verb*

We then created four experimental lists, each with a different random order. In each sentence, the TRY verb was replaced with a blank, and the six possible verbs were printed below it in alphabetical order.

### *3.1.2 Participants*

159 adult native speakers of Russian were recruited via e-mail announcements and through personal contacts. The participants were randomly assigned to one of the four lists. 25 participants did not supply responses for all verbs and were excluded. The data for the remaining 134 participants (28 males, 106 females) was entered into the analysis. The participants ranged in age from 17 to 64 (mean 30, SD = 10). The vast majority either held a university degree or was studying for one.

### *3.1.3 Procedure*

The participants were given the following instructions (in Russian):

You will be presented with 60 sentences from which a verb has been deleted.  
Read the sentences and the answer options and choose the verb that fits the context best from the list of 6 options. Work at a quick pace, don't think too

long over one answer, don't go back and change things: there are no right or wrong answers and we are interested in your first choice.

The experiment was administered online using Google Forms, and took about 15 minutes to complete.

To obtain comparable data from the model, we excluded the 60 test sentences from the Divjak (2010) dataset and trained the model on the remaining sentences. We then used the model to compute the probability for each of the six verbs in each of the test sentences.

## **3.2 Results and discussion**

### *3.2.1 Analysis 1: Model v. average participant*

In order to compare the model and the participants, we assumed that the model's response on the forced choice task would be the verb with the highest predicted probability for a given context. In the analysis that follows, we take the verb which actually occurred in the corpus to be the "correct" response. Of course the attested corpus example may be an unrepresentative one, so this is not necessarily the best way to evaluate the model. We will return to this issue in sections 3.2.2 and 5.

Since there were 60 sentences and 6 verbs, chance performance would be about 10/60; given the skewed distribution of verbs over experimental sentences discussed above, always choosing the same verb would result in a correct choice for between 4 and 20 out of 60 sentences, depending on the verb (see Table 2). Always selecting the TRY verb most frequently used in corpus data, *pytat'sja*, would have yielded a correct choice in 10 out of 60 sentences (see Table 2). The model predicted the verb that actually occurred in the corpus for 23 of the 60 test sentences – i.e.,

38% of the time. This is considerably lower than the performance on randomly chosen sentences (50% – see above), and reflects the fact that the testing set intentionally contained a larger proportion of verbs in highly ambiguous, or variable, contexts than would be the case in a random sample. The mean number of “correct” choices for the participants was 27.7, i.e., 46% of the time (SD 4.7) and the median was 28; the scores ranged from 13 to 38. Thus, there is considerable individual variation in humans (no doubt reflecting the fact that the participants often guessed), and the model performed about a standard deviation less well than the average human. In other words, although both model and speaker perform 2.5 to 3 times better than chance, they still make the “wrong” choice in more than half of all cases.

Tables 3, 4 and 5 provide summaries of the results by verb. Table 3 specifies the number of trigger sentences in which the verb that was used in the original corpus sentence was correctly retrieved by the model or by the human respondents. While the model performs particularly poorly on *pytat'sja* and *starat'sja*, the human respondents struggle with *silit'sja*, which is unsurprising as the verb is relatively infrequent (see Table 6 below for further discussion) and obsolescent.



Verb	Model (correct out of total)	Humans (correct out of total)
norovit'	7/12	10/12
probovat'	6/8	4/8
silit'sja	3/6	0/6
poryvat'sja	2/4	4/4
pytat'sja	1/10	8/10
starat'sja	4/20	11/20

Table 3. Target responses by verb for model and humans

Table 4 summarizes the choices made by our respondents for each verb. Each row in the table summarizes the participants' responses to sentences containing one of the six verbs. The numbers across the diagonal provide information about "correct" responses, i.e., proportion of times when participants supplied the verb that actually occurred in the corpus (e.g. 58% of the time for *norovit'*); the other figures in the same row give us the proportion of the time that other verbs were used in the same contexts. Thus, row 1 tells us that, on average, for corpus sentences that originally contained *norovit'*, participants supplied that verb 58% of the time, *probovat'* 3% of the time, *silit'sja* 5% of the time, and so on.

		Humans					
		<i>norovit'</i>	<i>probovat'</i>	<i>silit'sja</i>	<i>poryvat'sja</i>	<i>pytat'sja</i>	<i>starat'sja</i>
Corpus	<i>norovit'</i>	<b>0.58</b>	0.03	0.05	0.12	0.14	0.08
	<i>probovat'</i>	0.02	<b>0.57</b>	0.01	0.08	0.27	0.04
	<i>silit'sja</i>	0.03	0.04	0.15	0.02	<b>0.61</b>	0.15
	<i>poryvat'sja</i>	0.04	0.06	0.10	<b>0.57</b>	0.22	0.01
	<i>pytat'sja</i>	0.03	0.17	0.09	0.04	<b>0.50</b>	0.18
	<i>starat'sja</i>	0.03	0.07	0.10	0.01	0.39	<b>0.40</b>

Table 4. Results by verb: Human choices across sentences

It is clear from the table that the participants used all the verbs in each type of context, although they also had a strong preference for one of the verbs (and in the

case of *starat'sja* contexts, for two verbs, *pytat'sja* and *starat'sja*). Moreover, as we can see the highest values (in boldface) are not always on the diagonal. The verb *silit'sja* for example, was frequently replaced with *pytat'sja* by native speakers, and *starat'sja* is nearly equally often predicted as *pytat'sja* than as *starat'sja*.

For ease of comparison, we present the results for the model in a format similar to Table 4, containing the data provided by the respondents. Yet, it must be borne in mind that the 60 sentences were selected so as to contain a substantial number of cases with inherent variability, allowing virtually all of the 6 TRY verbs. Therefore, the average probabilities mask a substantial amount of variability in the sentence-wise verb-specific probability estimates. Thus, the first row in Table 5 gives us the predicted probability of *norovit'* in *norovit'* contexts (averaged across all sentences with *norovit'*), followed by the probabilities for the other verbs in *norovit'* contexts. Here the highest values are not always on the diagonal either, and *pytat'sja* and *starat'sja* as well as *poryvat'sja* are often replaced with *silit'sja* by the model.

		Model					
		<i>norovit'</i>	<i>probovat'</i>	<i>silit'sja</i>	<i>poryvat'sja</i>	<i>pytat'sja</i>	<i>starat'sja</i>
Corpus	<i>norovit'</i>	<b>0.47</b>	0.04	0.11	0.06	0.22	0.10
	<i>probovat'</i>	0.07	<b>0.55</b>	0.09	0.03	0.15	0.12
	<i>silit'sja</i>	0.19	0.01	<b>0.44</b>	0.03	0.14	0.19
	<i>poryvat'sja</i>	<b>0.33</b>	0.04	0.04	0.31	0.11	0.17
	<i>pytat'sja</i>	0.10	0.19	<b>0.27</b>	0.06	0.23	0.14
	<i>starat'sja</i>	0.07	0.16	<b>0.35</b>	0.04	0.16	0.22

Table 5. Results by verb: Model predictions across sentences

As we can see, the results for the model and the participants are broadly similar, but there are also some differences. The model has particular problems with

*starat'sja* and especially *pytat'sja*. The average predicted probability of *pytat'sja* in relevant contexts is 0.23 (Table 5), yet in the 60-sentence test sample, the model chose it as the most probable option (by a very narrow margin) in only one out of ten contexts in which *pytat'sja* was expected (Table 3). The corresponding figures for human participants, on the other hand, are considerably higher: the average predicted probability of *pytat'sja* in relevant contexts is 0.50 (Table 4) and humans chose it in 8/10 cases (Table 3). Furthermore, as can be seen by looking at the figures in column 5 of Table 4, participants often overgeneralized *pytat'sja*, using it in contexts where other verbs occurred in the corpus: in fact, for 20 out of the 50 sentences with verbs other than *pytat'sja*, the majority of the participants chose *pytat'sja*; the model did this much less frequently (in only 8 out of 50 cases). In contrast, the human participants struggled with the verb *silit'sja*, while the model did quite well with this verb.

These differences are likely to be due to frequency effects. As shown in Table 6 that contains the frequencies with which the TRY verbs appear followed by an infinitive, the verbs differ considerably in their frequencies: *pytat'sja* is the most frequent verb by a large margin, while *silit'sja* is one of the least frequent and is, in fact becoming obsolete.

Verb	Tokens in RNC	Relative frequency
<i>norovit'</i>	1266	0.02
<i>probovat'</i>	4023	0.07
<i>silit'sja</i>	492	0.01
<i>poryvat'sja</i>	241	<0.01
<i>pytat'sja</i>	32550	0.56
<i>starat'sja</i>	20011	0.34

Table 6. *Frequencies of the verbs followed by an infinitive in the Russian National Corpus (1992-2013)*

We know that humans are highly sensitive to frequency information (Ellis 2002, Divjak and Caldwell-Harris 2015), so it is not surprising that they tended to select the most frequent (and hence most general) verbs when they had no strong preference for a verb with a more specific meaning, i.e., when the contextual factors were not strong enough to clearly favour one outcome. This is especially the case in an experimental setting with only a small number of contexts, which limits the possibility of the effect of the estimated probabilities to emerge; (estimated) probabilities show their effect in the long run, and this typically requires more than a few dozen sentences. The model, in contrast, makes its predictions entirely on the basis of how often the sentence-wise combination of the variables discussed earlier (Table 1) is associated with each verb, as it had no access to information about the token frequencies of individual verbs (recall that the frequencies in the sample used for training were roughly equal by design to level the playing field for the contextual properties of interest). Moreover, the model considers relative frequencies of the outcome verbs, given the particular contexts, not overall proportions in general language usage.

To accommodate frequency information, we multiplied the predictions of the original model by the square root of each verb's relative frequency. Using the square root is a common practice when dealing with skewed distributions (Field, Miles and Field 2012); it is also psychologically realistic in that frequency effects are most noticeable at lower frequencies. Table 7 presents a summary of predictions for each verb; for ease of comparison with Tables 3 and 4, the figures given in Table 7 were converted to probabilities by dividing the frequency adjusted values for each verb in each sentence by the sum of the frequency adjusted values for all six verbs.

		Model					
		<i>norovit'</i>	<i>probovat'</i>	<i>silit'sja</i>	<i>poryvat'sja</i>	<i>pytat'sja</i>	<i>starat'sja</i>
Corpus	<i>norovit'</i>	0.27	0.04	0.03	0.01	<b>0.47</b>	0.17
	<i>probovat'</i>	0.05	<b>0.40</b>	0.04	0.01	0.31	0.19
	<i>silit'sja</i>	0.10	0.01	0.17	0.01	<b>0.36</b>	<b>0.35</b>
	<i>poryvat'sja</i>	0.20	0.04	0.02	0.08	<b>0.34</b>	<b>0.33</b>
	<i>pytat'sja</i>	0.05	0.12	0.10	0.01	<b>0.49</b>	0.23
	<i>starat'sja</i>	0.03	0.12	0.12	0.01	<b>0.36</b>	<b>0.36</b>

Table 7. Results by verb: Model predictions adjusted for frequency

The frequency-adjusted model predicted the target verb correctly in 28 of the 60 sentences – in other words, overall, it performed at exactly the same level as the average human participant. As expected, the frequency adjustment made the performance more human-like on *pytat'sja* and *starat'sja*. Moreover, like human participants, the frequency-adjusted model tended to overgeneralize *pytat'sja*, which is now the most frequently chosen option for all verbs except *probovat'*. It also undergeneralizes *silit'sja* and instead predicts it to be *pytat'sja* or *starat'sja*. On

the other hand, it performed less well than both original model and human participants on sentences with *norovit'* and *poryvat'sja*.

Thus, adding frequency information improved performance, but the overall improvement was relatively modest, and performance on some verbs actually deteriorated. This signals that the trade-off between frequency information and contextual information with which native speakers operate is more sophisticated than we can capture with a logistic regression model that runs on contextual features enriched with the frequency of the TRY verb in the targeted syntactic context.

### *3.2.2 Analysis 2: Model v. participants as a group*

All the analyses so far assumed that the verb which actually occurred in the corpus was the “correct” response. This is the fairest way to compare the model's performance to that of humans, but it is problematic in the sense that not all corpus examples are necessarily representative. In fact, since the corpus includes a high proportion of literary texts, it is possible that a number of the uses involved the author deliberately using an unusual verb for special effect. To determine whether this is the case, we conducted a second analysis to see how often the participants, the model, and the corpus “agreed” (i.e., both participants and the model choose the verb that occurred in the corpus) and how often they “disagreed”. For this analysis, the verb that was selected by the largest number of participants was deemed to be preferred: in other words, we treated each individual response as a “vote” for a particular verb in a particular sentence, and the verb that got the most votes was the winner. Logically, there are five possibilities:

- (1) the corpus, the model and the participants all agree;
- (2) the model chose the verb that occurred in the corpus while the participants prefer a different verb;
- (3) the participants prefer the verb that occurred in the corpus while the model prefers a different verb;
- (4) the model and the participants both prefer the same verb, but not the one that occurred in the corpus;
- (5) the model and the participants prefer different verbs, and the corpus contains yet another verb.

The results of the analysis are summarized in Table 8.

Type	Corpus	Model	Participants	Original model	Frequency adjusted model	% in frequency adjusted model
1	verb 1	verb 1	verb 1	17	19	32
2	verb 1	verb 1	verb 2	7	9	15
3	verb 1	verb 2	verb 1	18	16	27
4	verb 1	verb 2	verb 2	2	9	15
5	verb 1	verb 2	verb 3	16	7	12

Table 8. *Agreement between the corpus, the model and human participants*

As we can see, experimental items where the model and the participants agreed on a verb different from the verb used in the corpus account for 9 out of 60, i.e., 15% of all cases. In such cases, the choice of the verb attested in the corpus is arguably unusual or has become obsolete, and the verb preferred by the participants (and the model) should be regarded as (currently) “correct”. Thus, the accuracy figures given in the preceding section underestimate the participants' (and the model's) true performance by about 15%.

The corpus and the frequency-adjusted model agreed on 28 (19+9) out of the 60 sentences, that is in 46.6% of all cases. This is virtually identical to the average human performance: as indicated earlier, the mean human score was 27.7 and the median 28. However, as shown in the table, the humans *as a group* did considerably better, choosing the “correct” verb in 35 (19+16) or 58.3% of sentences.

Why should there be such a discrepancy between individual and group performance? One possibility is that the difference is due simply to the fact that, between them, 134 participants have experienced more verb tokens in relevant contexts than any one participant, and hence had more opportunities for learning the differences between the contexts (in the widest sense of the word, i.e. not necessarily limited to sentential contexts, and including subtle pragmatic differences and attitudes) in which the verbs occur. If this is the case, then we would expect older participants (who have had more experience, possibly including more experience with the type of texts the corpus contained) to perform better than younger participants. In order to test this possibility, we computed a Pearson product-moment correlation between participants' age and their scores in the experiment. The relationship turned out to be insignificant ( $r = -0.09$ ,  $p = 0.323$ ), suggesting that all participants have had enough relevant experience. Hence, it is unlikely that the difference between individual and group scores can be explained by the amount of experience – although it is possible, of course, that what matters more than sheer amount is the *type* of experience, for instance, exposure to particular genres.



Another possibility is that different participants relied on different features, and hence collectively the entire group were able to take advantage of more information than any one individual. This possibility is explored in Study 2.

### *3.2.3 Analysis 3: Using forced-choice responses as the test corpus*

In the last two sections, we compared human participants and the model by giving them both the same task: predicting the verb that actually occurred in each sentence in the test corpus. An alternative way to evaluate the model is to see how well it can predict the participants' responses: in other words, we can take all of the participants' responses (134 x 60 sentences) and used them as another test corpus for the model. In this section, we assess the model's performance on this test corpus.

A polytomous mixed-effect regression model (with participant as the random effect) of the type described in Section 2 achieves a likelihood-based pseudo-variance of MacFadden's  $R_L^2 = 26.2\%$  in explaining the individual categorical choices in the forced choice data using exactly the same model specification, i.e. variable combinations, as was used to explain the literary corpus data. This is slightly less than the original corpus-based model that explained 31% of the variation.

We can gain a better understanding of how the predictions for the forced choices corpus compare to the predictions for the literary texts corpus by inspecting the resulting odds tables. The odds from the forced choices model are represented in Table 9. Boldfaced odds greater than 1 signal variable levels in favour of a specific verb, odds less than 1 capture variable levels against a specific verb, and odds in parentheses denote insignificant variable levels. Take for example the fact that the TRY

verb occurs in a main clause (CLAUSE.MAIN.TRUE) which has significant positive odds in favor of *probovat'*, *pytat'sja*, *starat'sja* and *silit'sja* but neutral odds for *norovit'* and *poryvat'sja*. The comparatively high odds of a perfective aspect on the TRY verb (FINITE.ASPECT\_PERFECTIVE) in favor of *probovat'* 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.

Property/Verb	Probovat'	Pytat'sja	Starat'sja	Silit'sja	Norovit'	Poryvat'sja
CLAUSE.MAIN	<b>5.7</b>	<b>1.578</b>	<b>1.425</b>	<b>1.526</b>	0.4138	0.5319
FINITE.ASPECT_PERFECTIVE	<b>3.72</b>	-0.9	(1.033)	0.1199	0.3545	0.3267
FINITE.MOOD_GERUND	<b>2.449</b>	<b>2.318</b>	<b>1.014</b>	<b>2.419</b>	0.2614	(0.6499)
FINITE.MOOD_INDICATIVE	0.5114	(0.909)	0.6188	(0.8366)	<b>2.121</b>	(1.072)
FINITE.TENSE_PAST	(1.059)	(1.1)	<b>1.422</b>	<b>1.342</b>	(1.002)	(1.101)
INFINITIVE.ASPECT_IMPERFECTIVE	(0.9793)	0.4584	<b>2.54</b>	0.3815	0.05022	(1.051)
INFINITIVE.CONTROL_HIGH	0.3666	0.4629	0.3126	0.2984	<b>7.784</b>	(1.219)
INFINITIVE.SEM_COMMUNICATION	(0.9066)	0.6828	0.8294	0.3263	0.04814	(0.907)
INFINITIVE.SEM_METAPHORICAL_MOTION	(1.26)	0.3235	0.5285	(0.9971)	0.7401	<b>2.356</b>
INFINITIVE.SEM_METAPHORICAL_PHYSICAL_EXCHANGE	0.2375	0.5212	(1.098)	0.06105	<b>2.098</b>	(1.159)
INFINITIVE.SEM_MOTION	0.4963	0.3443	0.2656	(0.9185)	<b>1.278</b>	<b>8.544</b>
INFINITIVE.SEM_MOTION_OTHER	0.2514	0.3929	(1.332)	0.431	(1.3)	(0.8574)
INFINITIVE.SEM_PHYSICAL	0.1032	0.1434	<b>1.561</b>	0.1703	<b>5.231</b>	(0.7268)
INFINITIVE.SEM_PHYSICAL_OTHER	(0.8627)	0.4187	0.2042	0.4599	(1.048)	<b>1.894</b>
SENTENCE.DECLARATIVE	<b>2.104</b>	<b>1.946</b>	<b>7.562</b>	<b>2.369</b>	0.2356	<b>2.209</b>
SUBJECT.SEM_ANIMATE_HUMAN	<b>1.916</b>	<b>1.704</b>	<b>2.395</b>	<b>1.632</b>	0.2488	(1.199)

Table 9. *Verb specific odds per property for all six Russian verbs in the forced-choice corpus*

Even at first glance, it is clear that while some verbs have clearly different profiles, others are more similar to each other. For example, *probovat'*, *pytat'sja*, *silit'sja* and *starat'sja* share four of their favourable odds and the differences between *probovat'* and *pytat'sja* in terms of odds in favour are marginal (perfective aspect triggers *probovat'*). Other verbs, such as *norovit'* and *poryvat'sja* are markedly differ-

ent, sharing at most one favoured property with the four other verbs. As was the case with the odds derived from the literary corpus data (presented in Table 10 below), overall, the presence of the infinitive plays a significant role in the selection of *norovit'* and *poryvat'sja*, but for the other four verbs it is either much less relevant or even signals repulsion in the case of *pytat'sja*.

Property/Verb	Probovat'	Pytat'sja	Starat'sja	Silit'sja	Norovit'	Poryvat'sja
(Intercept)	1:22	1:12	1:47	(1:5.8)	(1:2.2)	1:3380
CLAUSE.MAIN	<b>3.4:1</b>	1:1.6	(1:1.1)	(1:1)	(1:1.2)	(1:1)
FINITE.ASPECT_PERFECTIVE	<b>29:1</b>	(1.1:1)	(1.1)	(1:4.9e <sup>7</sup> )	(1:1,8e <sup>8</sup> )	(1:3,0e <sup>7</sup> )
FINITE.MOOD_GERUND	1:8.3	(1.2:1)	<b>2.2:1</b>	<b>7:1</b>	1:6	(2.8:1)
FINITE.MOOD_INDICATIVE	1:2.8	(1.3:1)	(1.9:1)	(2.1)	(1:1.2)	(1.8:1)
FINITE.TENSE_PAST	(1:1)	<b>2.4:1</b>	1:2	<b>2.1:1</b>	1:3.3	3.3:1
INFINITIVE.ASPECT_IMPERFECTIVE	<b>6.1:1</b>	1:2.7	<b>4:1</b>	1:10	1:2.9	(1:1)
INFINITIVE.CONTROL_HIGH	(1:1.2)	<b>3.1:1</b>	<b>1.6:1</b>	1:6.4	<b>2.6:1</b>	<b>4.7:1</b>
INFINITIVE.SEM_COMMUNICATION	<b>2.1:1</b>	1:1.9	(1:1.6)	(1:1)	(1.2:1)	<b>8.4:1</b>
INFINITIVE.SEM_EXCHANGE	(1.4:1)	(1:1.9)	(1:1.5)	1:11	<b>7.7:1</b>	<b>9.1:1</b>
INFINITIVE.SEM_METAPH..._MOTION	(1.5:1)	(1:1)	(1:1.5)	1:3.7	<b>6.1:1</b>	(1.9:1)
INF....SEM_METAPH..._PHYS..._EXCH...	(1:1.3)	1:2.6	(1.8:1)	1:3	<b>4:1</b>	(4:1)
INF....SEM_METAPH..._PHYS..._OTHER	(1.3:1)	(1:1.3)	(1:1.1)	(1:1.3)	<b>2.7:1</b>	(1.3:1)
INFINITIVE.SEM_MOTION	(1.7:1)	1:4.2	1:3.2	1:4.5	<b>8.1:1</b>	<b>19:1</b>
INFINITIVE.SEM_MOTION_OTHER	(2.6:1)	(1:1.5)	1:3.6	(1:1.3)	<b>4.5:1</b>	<b>5.1:1</b>
INFINITIVE.SEM_PHYSICAL	<b>3.9:1</b>	<b>1.4:1</b>	(1:1.8)	(1:1.1)	<b>6:1</b>	(1.6:1)
INFINITIVE.SEM_PHYSICAL_OTHER	<b>2.5:1</b>	(1:1.5)	1:2.1	1:2.6	<b>6.1:1</b>	<b>3.1:1</b>
SENTENCE.DECLARATIVE	1:2.8	(1:1.1)	<b>2.8:1</b>	(3.2:1)	(1:1)	(1.3:1)
SUBJECT.SEM_ANIMATE_HUMAN	(1.5:1)	(1.4:1)	<b>2.5:1</b>	(1:1.1)	1:4	<b>4.1:1</b>

Table 10. Verb specific odds per property for all six Russian verbs in the original test corpus

When we compare the odds tables (Table (9) and Table (10)) in more detail, we see that the odds in favour of one and the same verb are different depending on the corpus. The verb that shows least variation in this respect is *starat'sja* that is in both datasets likely to be used in declarative sentences with a human subject, if the TRY verb occurs in a gerund or if the infinitive has imperfective aspect marking; in the literary corpus data high control over the infinitive action was another trigger, while in the forced choices corpus data occurring in a main clause, to describe a past attempt at a physical action turned out to be triggers. Other verbs, such as *pytat'sja* and *poryvat'sja* seem to be triggered by entirely different sets of properties in the literary

data and the forced choices data. *Pytat'sja* is such an example: while in the literary corpus model variables levels such as past tense, a high level of control over the infinitive action and physical activities trigger *pytat'sja*, in the forced choices model it is a human subject, occurrence as gerund, being used in a main clause and in a declarative sentence that trigger the verb. Furthermore, in the literary corpus model, 9 out of the 19 variables are insignificant, while in the forced choices model only 3 out of 19 are insignificant and 9 out of 19 are significantly against. Nevertheless the aggregated effects, i.e. overall, the correlation between the corpus-based probabilities and the forced-choice proportions stands at 0.46 ( $t = 9.8012$ ,  $df = 358$ ,  $p < 0.001$ ).

Why then are the odds in the models different? Primarily, because the selection of the sample sentences in the forced choices model and the frequencies of the properties associated with these sentences is different. The sample of 60 sentences presented to the subjects is much more limited in terms of the range of possible contextual properties and property combinations that it contains than the literary corpus; this affects the contribution each property makes to the choice of one option over another. This key difference also has to be borne in mind when attempting to use elicited data to investigate the behavior of properties for which no or not enough corpus data is available (Bresnan 2007).

#### **4. STUDY 2: MODELLING GROUP EFFECTS AND INDIVIDUAL DIFFERENCES**

We know that language learners are highly sensitive to frequency. However, they cannot track the frequency of everything they encounter – so how do they know what to track? This problem has led many researchers (see, for example, Golinkoff, Mervis and Hirsh-Pasek 1994, Markman 1987, Woodward and Markman 1998) to

conclude that humans have innate biases which lead them to focus on some features while ignoring others. Note that this conclusion is based on an implicit assumption, namely, that all speakers of the same language converge on (more or less) the same grammar. While this assumption is quite widespread, there is now considerable evidence that it is incorrect: there are in fact significant differences between individual speakers' grammars (see Dąbrowska 2012 and Dąbrowska 2015 for reviews). It is possible, then, that different individuals concentrate on different features, and this could explain why the group did better than the average individual: between them, they are able to cover all the relevant features. We explore this possibility in the second study.

#### **4.1 Method**

As explained earlier, the *try* verb dataset was coded for 87 features, but the model developed by Divjak and Arppe (2013) included only 18 hand-picked features. Comparative modelling suggested that different variable combinations could achieve comparable results, and that omitting some affected prediction accuracy more than omitting others. Here, we take this line of thought further and apply it in modeling individual differences. In total 134 different “dumb” models were constructed to match the number of participants in study 1. Each of these “dumb” models used a different, randomly selected subset of 18 variables. These variables were chosen from the 25 that were retained after complying with the requirements for gracefully fitting the individual one-vs-rest models, i.e. the variables occur at least once in the Forced Choice stimuli, and at least once with all of the 6 TRY verbs in the full dataset. Two exactly collinear properties were excluded (dealing with the complementary

aspect marking of the finite and infinitive verbs); other than that, collinearity was not considered since it does not affect overall prediction accuracy if it is pervasive, i.e. present not just in the sample but throughout the population (cf. Harrell 2001: 65), which is what we assume here. As before, the verb with the highest predicted probability was regarded as the model's choice and the scores were compared to the scores we got from the respondents.

#### 4. 2. Results and discussion

The prediction accuracy of these 134 models ranges from 30% to 45% (mean 39% and median 38.3%). For robustness, we also ran this same procedure using 2500 random models; this yielded a wider range but similar mean and median. The results are presented in the Appendix.

The worst and best models share 12 out of 18 contextual variables, as illustrated in Table 11. The table shows that certain properties such as (present) tense and (imperfective) aspect of the finite verb, as well as aspect of the infinitive contribute to the individual profiles of the verbs. Although tracking these significantly improves prediction accuracy, they are not typically included in lexical semantic studies.

<b>Worst model</b>	<b>Best model</b>
CLAUSE.MAIN	
	FINITE.ASPECT_IMPERFECTIVE
FINITE.MOOD_GERUND	
FINITE.MOOD_INDICATIVE	FINITE.MOOD_INDICATIVE
	FINITE.TENSE_PRESENT
	INFINITIVE.ASPECT_IMPERFECTIVE
	INFINITIVE.ASPECT_PERFECTIVE
INFINITIVE.CONTROL_HIGH	INFINITIVE.CONTROL_HIGH
INFINITIVE.CONTROL_MEDIUM	INFINITIVE.CONTROL_MEDIUM
INFINITIVE.SEM_COMMUNICATION	INFINITIVE.SEM_COMMUNICATION

	INFINITIVE.SEM_METAPHORICAL_PHYSICAL_EXCHANGE
INFINITIVE.SEM_METAPHORICAL_MOTION	
INFINITIVE.SEM_METAPHORICAL_MOTION_OTHER	
	INFINITIVE.SEM_MOTION
INFINITIVE.SEM_MOTION_OTHER	INFINITIVE.SEM_MOTION_OTHER
INFINITIVE.SEM_PERCEPTION	
INFINITIVE.SEM_PHYSICAL	INFINITIVE.SEM_PHYSICAL
INFINITIVE.SEM_PHYSICAL_OTHER	INFINITIVE.SEM_PHYSICAL_OTHER
SENTENCE.DECLARATIVE	SENTENCE.DECLARATIVE
SENTENCE.EXCLAMATIVE	SENTENCE.EXCLAMATIVE
SENTENCE.NONDECLARATIVE	
SUBJECT.SEM_ANIMATE_ANIMAL	SUBJECT.SEM_ANIMATE_ANIMAL
SUBJECT.SEM_ANIMATE_HUMAN	SUBJECT.SEM_ANIMATE_HUMAN
SUBJECT.SEM_INANIMATE_MANMADE	SUBJECT.SEM_INANIMATE_MANMADE

Table 11. *Properties used in the best and worst models out of 134 random models*

The models supplied the target verb (that is to say, the verb that actually occurred in the corpus) on average in 23.4 out of the 60 sentences, i.e. 39% of the time (median 23, SD 1.7, range 18-27). Thus, the average level of accuracy of the “dumb” models was virtually identical to that of the hand-crafted model, which, as we have seen, selected the target verb for 23/60 sentences, and slightly below that of human participants who scored 28/60. Interestingly, however, there was much less variation in the “dumb” models' accuracy scores than in humans: recall that the standard deviation for humans was 4.7 – almost three times larger than for the dumb models, and the range of scores was 13-38 – almost four times larger. This is rather surprising, and suggests that it does not really matter which contextual features humans track, as long as they track enough features.<sup>5</sup>

<sup>5</sup> How many features would be enough requires further investigation, but preliminary results from a 1000-fold random selection of 18 variables from the original full 26-variable set, as reported in Divjak & Arppe (2013), reveal the following: the mean accuracy for these 1000 random models was 45.95%, ranging from 26.87% to 51.59. The best 100 random Russian models (with accuracy values ranging from 49.44% to 51.59%) had on average 11 (60.0%) variables values in common with each other, ranging from as few as 6 up to as many as 15 common variables in individual pairwise comparisons. Moreover, the best and worst models had only 8 variables (44%) in common, which probably explains the substantial difference in model performance. We do not pursue this question further here

Allowing the models to “vote” in the same way as the human participants in Study 1 resulted in a negligible improvement in performance, from 23.4 to 24. This is probably due to the fact that, in contrast to the humans, the models' property space remained constrained: although the 134 models were able to track more properties, only 25 out of the 87 annotated for were available to them (Divjak 2010). The improvement in performance that we observed for the human participants strongly suggests that they not only must have tracked different property constellations, but that they had access to a larger range of properties than were considered in our study. We return to this issue in the concluding section.

### **5. STUDY 3 – ACCEPTABILITY RATINGS**

In a third study, we compare the probability the corpus model assigns to encountering each of the 6 verbs in the 60 test sentences to the acceptability ratings that adult native speakers would assign to those combinations. Several papers have investigated the relation between corpus-based frequencies and native speaker judgments (Featherston 2005; Kempen & Harbusch 2005; Arppe & Järviö 2007; Klavan 2012; Bermel & Knittl 2012), including the relation between probabilities conditioned on one contextual element and acceptability ratings (Divjak 2008; in press). This study is, however, the first to correlate corpus-based probabilities for the choice of one verb over another, conditioned on *all* other elements present in the sentence, with native speaker ratings of the suitability of these verbs in those

---

because our interest is not in finding the most parsimonious model, but rather in exploring the impact of the contextual effects that we had selected on the basis of prior studies. As Tarpey (2009), echoing Box (1979), put it, “in any given data analysis situation, a multitude of models can be proposed. Most of these will be useless ... and perhaps a few will be useful.”



sentences. It therefore measures, in more detail than the forced choice task, how well model-predictions align with native speaker intuitions.

## **5.1 Method**

### ***5.1.1 Materials***

The same 60 sentences as selected for the forced choice task presented in Section 2.1.1 were used in the acceptability ratings task. Yet, instead of offering them to native speakers in the form in which they occurred in the corpus, we created six versions of each sentence, using each of the six TRY verbs. Six stimulus sets were derived in such a way that the probability distributions estimated by the polytomous logistic regression model were equally well represented across all 6 sets. Within each set, the sentence order was randomized once, to avoid having more likely or more unlikely items cluster together, and each participant saw 10 cases of each verb. Since the literary corpus model predicted probabilities for all 6 verbs in each context, precise predictions about acceptability are available for all possible verb-by-context combinations and this in the form of probabilities of occurrence.

### ***5.1.2 Participants***

103 adult native speakers of Russian were recruited via e-mail announcements and through personal contacts. The participants were randomly assigned to one of the six lists. The vast majority either held a university degree or was studying for one. Respondents could enter a prize draw where in total 6 Amazon or Ozon vouchers of £15 each could be won.

### 5.1.3 Procedure

The participants were given the following instructions (in Russian):

In this experiment you will be asked to rate how natural sentences sound. We are specifically interested in the use of verbs meaning TRY such as *probovat'*, *pytat'sja*, *starat'sja*, *silit'sja*, *norovyt'* and *poryvat'sja*. There are 72 sentences in total and we would like you to rate them on a scale from 1 (sounds very strange) to 10 (sounds completely natural). Work at a quick pace, don't think too long over one answer, don't go back and change things: there are no right or wrong answers and we are interested in your first choice.

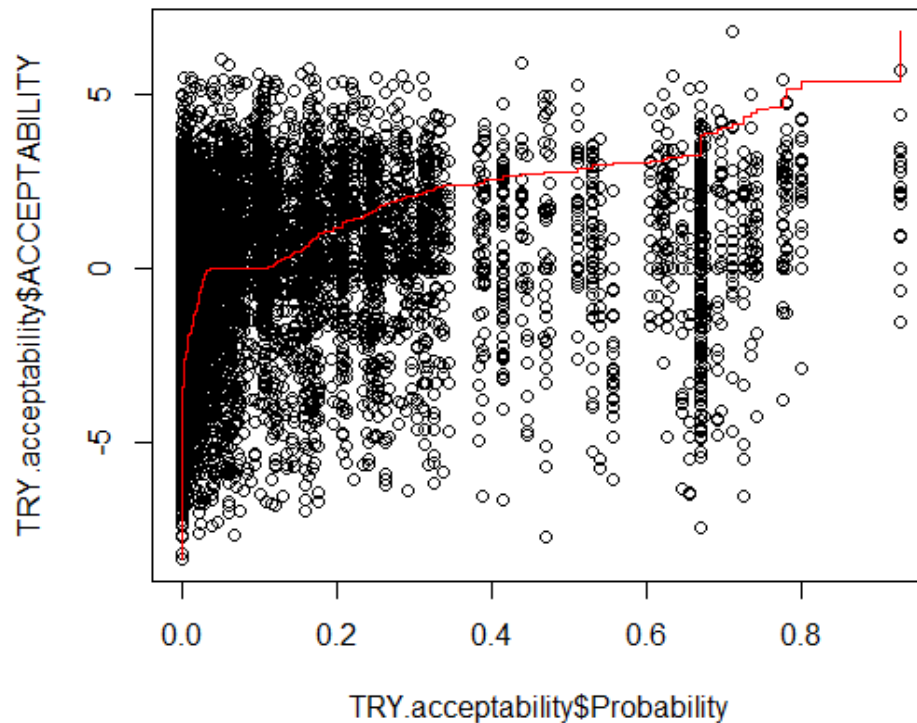
The experiment was administered online using Google Forms, and took about 15 minutes to complete.

## 5.2 Results and discussion

For the analysis of the data, the raw acceptability ratings were residualized against participant and position of the sentence in the experiment, so that what remained of their rating was free of differences in how participants used the scale, or how their ratings would change over the course of the experiment. The residualized ratings were rescaled so that the ratings per participants used the entire range 1-10. Our results, visualized in Figure (2), show a clear two-way distinction between low-probability items for which the acceptability can vary, with acceptability then converging and finally linearly increasing from  $p$  values of 0.15, as shown by the grey line. This means that, whereas the high probability of a verb given its context by and large entails acceptability of that verb in that context (as witnessed by the relatively empty lower right hand quadrant), the (relative) low probability of a verb given its

context does not entail lower acceptability. In other words, the probability that the corpus model calculates for encountering each of the 6 verbs in the 60 test sentences correctly predicts how acceptable each sentence will be for adult native speakers, but only for combinations that the model considers likely.

Figure (2): Residualized and rescaled acceptability ratings plotted against probability



This result confirms previous findings by Arppe & Järviö (2007), Divjak (2008) and Bermel & Knittl (2012) who concluded that meta-linguistic acceptability relates to probability in a non- straightforward way, as both high and low probable items may exhibit a high degree of acceptability. In fact, Figure (2) shows that the low probability of an item given its context can correlate with any degree of acceptability.

### 3. Conclusion

The goal of much computational modelling work is to develop the best – i.e., most accurate – model of the phenomenon in question. As we have seen, once its predictions were adjusted for verb frequency, the Divjak and Arppe (2013) model for choosing between 6 Russian near-synonyms was able to predict the verb that actually occurred in the test corpus with 47% accuracy. While this may seem disappointing at first, a comparison with the performance of 134 human judges reveals that this is actually an excellent result. Many linguistic phenomena are simply not fully predictable, and if are interested in modelling *human* knowledge, we should compare our models' performance to that of human respondents.

To investigate this further, we created 134 models which used a random selection of the features, and they all performed within the human range. This demonstrates that a very large number of models can approximate human behaviour, which is in itself hugely varied. Divjak & Arppe (2013) noted already that “there would appear to exist some redundancy among the properties, which testifies to the inherent multicollinearity of linguistic variables that is extremely difficult, if not impossible, to eliminate, as well as to a degree of potentially significant divergence in possible property combinations leading to similar model fit and accuracy”. Any given feature seems predictable from many other features. Because of this redundancy, an utterance can be produced in (unobservably) different ways, which explains how individual differences and uniformity across the community can co-exist (Barth & Kapatsinski 2014, Dąbrowska 2013, Dąbrowska 2014, Hurford 2000). Thus, while multicollinearity can be a major headache for statistical modelling (but see Harrell 2001), it may be a blessing for language learners, in that it enables

speakers to behave in a way that is broadly similar to that of other speakers even when they all have different underlying grammars. This, combined with the considerable differences in the performance of *human* participants, suggests that rather than trying to find the single “best” model, it may be more productive to develop a range of models modelling the range of human performance (as already suggested by Lauri Carlson, cf. Arppe 2008: 208); for a practical implementation, see Barth & Kapatsinski 2014).

Second, our results suggest improvements to future models of linguistic data. Study 3 confirms that meta-linguistic acceptability relates to probability in a non-straightforward way, as likely combinations tend to be judged as acceptable but low probable items may exhibit a high as well as a low degree of acceptability. In the case of a 6-way choice, the absence of a clear correlation between probability and acceptability is likely due to the fact that low probability can be the result of competition between a number of equiprobable items, i.e. items that are equally likely given the context. This indicates that such information would need to be brought into linguistic models to increase their cognitive reality. Although the difference between the onomasiological and semasiological components of word meaning dates back to Structuralism and Geeraerts et al. (1994) have outlined an overall framework for quantitative onomasiology, we are not aware of any corpus-based modelling work that would have factored onomasiology into its statistical model. Efforts are underway (Author 1 et al., in progress) to model this phenomenon using measures from Information Theory such as entropy that captures uncertainty.

Finally, the results reported here also raise some new questions. Although the accuracy of the frequency-adjusted Divjak and Arppe (2013) model was similar to that of the average Russian speaker, it did not perform as well as the participants as a group. We hinted earlier that this is probably due to the fact that the individual differences between speakers were much larger than those between the models used in Study 2. This suggests that speakers differ not just in which features they track, but also how many features they are able to track, and possibly also in their sensitivity to frequency effects. A second line of inquiry that will shed light on this issue is more technical in nature and considers alternative ways of evaluating the model's performance, by steering clear of considering the highest probability option as the chosen option (cf. the criticism levelled at measures of classification accuracy for multivariate models that model probability distributions, cf. Hosmer & Lemeshow 2000). And finally, if the field of linguistics adopts the approach advocated in this paper, and starts to test corpus-based models against human performance routinely, the cognitive plausibility of the algorithm should be considered as a goodness-of-fit criterion, particularly in research within cognitive linguistic paradigms. Baayen et al (2013) have shown that statistical classifiers based on cognitively realistic approximations of how humans learn such as NDL perform as well as regression models for binary choices. Preliminary results support this finding for more complex corpus models that predict a 4-way polytomous choice (Arppe & Baayen 2011).

Capitalizing on the findings we have presented will help us address some interesting theoretical questions that have hitherto remained unanswered. As noted earlier, language acquisition researchers worry about how learners know which

features to track. The results of Study 2 suggest that it does not really matter what exactly learners track, as long as they track enough features. The results of the random variable selection in particular point to overlapping property combinations making up the core of a lexeme; this would make it possible for speakers to draw largely similar interpretations regarding lexemes even though the individual properties they have tracked and recorded differ. What this implies for the degree to which all speakers of a language share the same contextual property associations, and thus also any abstract prototypes derived from such sets of properties, requires further research.

## References

Arppe, Antti. 2008. Univariate, bivariate and multivariate methods in corpus-based lexicography - a study of synonymy. Publications of the Department of General Linguistics, University of Helsinki, No. 44. [PhD Dissertation]. URN: <http://urn.fi/URN:ISBN:978-952-10-5175-3>

Antti Arppe. 2013a. 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). <http://cran.r-project.org/web/packages/polytomous/vignettes/exemplars2prototypes.pdf>

Arppe, Antti & Dana Abdulrahim. 2013. Converging linguistic evidence on two flavors of production: The synonymy of Arabic COME verbs. *Second Workshop on Arabic Corpus Linguistics*, University of Lancaster, UK, 22-26 July, 2013. <http://www.comp.leeds.ac.uk/eric/wacl/wacl2proceedings.pdf>

Arppe, Antti & R. Harald Baayen. 2011. Statistical classification and principles of human learning. *4th Conference on Quantitative Investigations in Theoretical Linguistics (QITL4)*, Humboldt-Universität zu Berlin, Germany, 28-31 March 2011.

Arppe, A. & J. 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. H., Janda, L. A., Nessel, T., Dickey, S., Endresen, A., Makarova, A. 2013. Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics* 37: 253-291.

Barth, D., & V. Kapatsinski. (2014 – in press). A multimodel inference approach to categorical variant choice: construction, priming and frequency effects on the choice between full and contracted forms of *am*, *are* and *is*. *Corpus Linguistics & Linguistic Theory*.

Bermel, N. & L. Knittl. 2012. Corpus frequency and acceptability judgements: A study of morphosyntactic variants in Czech, *Corpus Linguistics and Linguistic Theory*, 8(2), 241-275.

Box, G. E. P. (1979), "Robustness in the strategy of scientific model building", in Launer, R. L.; Wilkinson, G. N., *Robustness in Statistics*, Academic Press, pp. 201–236.

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

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*, 77-96. Berlin: Mouton de Gruyter. [Studies in Generative Grammar].



Bresnan, J, Cueni, A., Nikitina, T. & Baayen, R. H. 2007. Predicting the dative alternation. In G. Boume, I. Krämer, and J. Zwarts (Eds.), *Cognitive Foundations of Interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Dawes, Robyn M., Faust, David & Meehl, Paul E. 1989. Clinical versus actuarial judgment. *Science* 243: 1668-1644.

De Sutter, Gert, Dirk Speelman &, Dirk Geeraerts. 2008. Prosodic and syntactic-pragmatic mechanisms of grammatical variation: the impact of a postverbal constituent on the word order in Dutch clause final verb clusters. *International Journal of Corpus Linguistics* 13 (2): 194-224.

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

Divjak, D. 2004. *Degrees of verb integration: conceptualizing and categorizing events in Russian*. Unpublished PhD dissertation, KULeuven, Belgium.

Divjak, D. 2008. On (in)frequency and (un)acceptability. In: Barbara LewandowskaTomaszczyk (ed.). *Corpus Linguistics, Computer Tools and Applications - state of the art*. Frankfurt a. Main: Peter Lang, 213-233. [Łódź Studies in Language]

Divjak, D. 2010. *Structuring the Lexicon: a Clustered Model for Near-Synonymy*. Berlin: De Gruyter. [Cognitive Linguistics Research 43].

Divjak, D. (in press). The role of lexical frequency in the acceptability of syntactic variants: evidence from that-clauses in Polish. *Cognitive Science*.

Divjak, D. & A. Arppe. 2013. Extracting prototypes from exemplars. What can corpus data tell us about concept representation? *Cognitive Linguistics*, 24 (2): 221-274

Divjak, D. & C. Caldwell-Harris. 2015. Frequency and entrenchment. In Ewa Dąbrowska & Dagmar Divjak (eds.). *Handbook of Cognitive Linguistics*, 53-75. De Gruyter-Mouton. [HSK].

Divjak, D. & St. Th. Gries. 2006. Ways of Trying in Russian. Clustering Behavioral Profiles. *Journal of Corpus Linguistics and Linguistic Theory*, 2 (1): 23-60.

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

Dąbrowska, E. 2012. Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism*, 2, 219-253.

Dąbrowska, Ewa. 2013. Functional constraints, usage, and mental grammars: A study of speakers' intuitions about questions with long-distance dependencies. *Cognitive Linguistics* 24: 633 - 665.

Dąbrowska, Ewa. 2014. Recycling utterances: A speaker's guide to sentence processing. *Cognitive Linguistics* 25: 617-653.

Dąbrowska, E. (2015). Individual differences in grammatical knowledge. In E. Dąbrowska & D. Divjak (Eds.), *Handbook of Cognitive Linguistics* (pp. 649-667). Berlin De Gruyter Mouton.

Ellis, Nick C. 2002 Frequency effects in language acquisition: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24: 143-188.

Ellis, Nick C. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24: 143-188.

Featherston, Sam. 2005. The Decathlon Model. In Kepser, Stephan and Marga Reis (eds.), *Linguistic Evidence. Empirical, Theoretical and Computational Perspectives*, 187-208. Berlin/New York: Mouton de Gruyter. [Studies in Generative Grammar 85.]

Field, Andy, Miles, Jeremy and Field, Zoë. 2012. *Discovering Statistics Using R*. Los Angeles: Sage. Firth, J. R. 1957. *Papers in Linguistics 1934–1951*. London: Oxford University Press

Ford, M. 1983. A Method for Obtaining Measures of Local Parsing Complexity throughout Sentences. *Journal of Verbal Learning and Verbal Behavior*, 22:203-18.

Ford, Marilyn & Joan Bresnan. 2012. 'They whispered me the answer' in Australia and the US: A comparative experimental study. Forthcoming in *From Quirky Case to Representing Space: Papers in Honor of Annie Zaenen*, edited by Tracy Holloway King and Valeria de Paiva. Stanford: CSLI Publications.

Ford Marilyn & Joan Bresnan. 2013. Using convergent evidence from psycholinguistics and usage. In Manfred Krug and Julia Schlüter (eds.) *Research Methods in Language Variation and Change*, 295-312. Cambridge University Press.

Dirk Geeraerts, Stefan Grondelaers & Peter Bakema. 1994. *The Structure of Lexical Variation*. Berlin: Mouton de Gruyter.

Golinkoff, R., Mervis, C., & Hirsh-Pasek, K. 1994. Early object labels: The case for a developmental lexical principles framework. *Journal of Child Language*, 21, 125-156.

Gries, Stefan Th. 2002 Evidence in linguistics: Three approaches to genitives in English. In Brend, Ruth M., William J. Sullivan and Arle R. Lommel (eds.), *LACUS Forum XXVIII: What Constitutes Evidence in Linguistics?*, 17-31. Fullerton: LACUS.

Gries, Stefan Th. 2003. *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. New York / London: Continuum International Publishing Group.

Grondelaers, Stefan; Speelman, Dirk. 2007. A variationist account of constituent ordering in presentative sentences in Belgian Dutch. *Corpus Linguistics and Linguistic Theory* 3: 161-193.

Grove, William M., David H. Zald, Boyd S. Lebow, Beth E. Snitz, and Chad Nelson. 2000. Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment* 12: 19-30.

Harrell, Frank E. 2001. *Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression and Survival Analysis*. New York: Springer-Verlag.

Kempen, Gerard and Karin Harbusch. 2005. Grammaticality ratings and corpus frequencies. In Kepser, Stephan and Marga Reis (eds.), *Linguistic Evidence. Empirical, Theoretical and Computational Perspectives*. (Studies in Generative Grammar 85.) Berlin/New York: Mouton de Gruyter, 329-349.

Harris, Z. (1954). Distributional structure. *Word*, 10(23): 146-162.

Hurford, James R. 2000. Social transmission favours linguistic generalisation. In Chris Knight, Michael Studdert-Kennedy and James R. Hurford (eds.) *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, 324–352. Cambridge: Cambridge University Press.

Keller, Frank. 2000. Gradiance in grammar: Experimental and computational aspects of degrees of grammaticality. Unpublished . Ph.D thesis, University of Edinburgh.

Kempen, Gerard and Karin Harbusch. 2005. Grammaticality ratings and corpus frequencies. In Kepser, Stephan and Marga Reis (eds.). *Linguistic Evidence. Empirical, Theoretical and Computational Perspectives*, 329-349. Berlin/New York: Mouton de Gruyter. [Studies in Generative Grammar 85]

Kilgarriff, A. 2005. Language is never ever ever random. *Corpus Linguistics and Linguistic Theory* 1 (2): 263-276.

Klavan, J. 2012. Evidence in linguistics: corpus-linguistic and experimental methods for studying grammatical synonymy. *Dissertationes Linguisticae Universitatis Tartuensis* 15.

Markman, E. M. (1987). How children constrain the possible meanings of words. In U. Neisser (Ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization* (pp. 255-287). Cambridge: Cambridge University Press.

Penke, Martina and Anette Rosenbach 2004. What counts as evidence in linguistics: An introduction. *Studies in Language* 28(3): 480-526.

Roland, D., Elman, J.L., and Ferreira, V.S. 2006. Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition*, 98: 245–272.

Stanovich, Keith E. 2010. *How to Think Straight about Psychology*. 9th ed. Boston: Pearson.

Sorace, Antonella and Frank Keller. 2005. Gradiance in linguistic data. *Lingua* 115(1): 1497-1524.

Tarpey, Thaddeus. 2009. All Models are Right . . . Most Are Useless. Joint Statistical Meetings. Washington, DC.

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

Wasow, T. & Arnold, J. 2003. Post-verbal constituent ordering in English. In G. Rohdenburg and B. Mondorf (Eds,) *Determinants of Grammatical Variation in English*, 119–154. Berlin: Mouton de Gruyter.

Woodward, A. L., & Markman, E. M. (1998). Early word learning. In D. Kuhn & R. S. Siegler (Eds.), *Handbook of Child Psychology. Volume 2: Cognition, Perception and Language* (pp. 371-420). New York: John Wiley & Sons.

## Appendix

For robustness, we also ran this same procedure using 2500 random models. The prediction accuracy now ranges from 23.3% to 48.3% (but with comparable mean 38.6% and median 38.3% as the 134 models). These models also share 12 out of 18 properties, as illustrated in Table 10.

Table 10. *Properties used in the best and worst models out of 2500 random models*

Worst model	Best model
CLAUSE.MAIN	CLAUSE.MAIN
	FINITE.MOOD_GERUND
	FINITE.MOOD_INDICATIVE
	FINITE.TENSE_PAST
	FINITE.TENSE_PRESENT
INFINITIVE.ASPECT_IMPERFECTIVE	INFINITIVE.ASPECT_IMPERFECTIVE
INFINITIVE.ASPECT_PERFECTIVE	INFINITIVE.ASPECT_PERFECTIVE
INFINITIVE.CONTROL_HIGH	
INFINITIVE.CONTROL_MEDIUM	INFINITIVE.CONTROL_MEDIUM
INFINITIVE.SEM_COMMUNICATION	
INFINITIVE.SEM_METAPHORICAL_MOTION	
INFINITIVE.SEM_METAPHORICAL_MOTION_OTHER	
INFINITIVE.SEM_METAPHORICAL_PHYSICAL_EXCHANGE	
INFINITIVE.SEM_MOTION	INFINITIVE.SEM_MOTION
INFINITIVE.SEM_MOTION_OTHER	INFINITIVE.SEM_MOTION_OTHER
INFINITIVE.SEM_PERCEPTION	INFINITIVE.SEM_PERCEPTION
INFINITIVE.SEM_PHYSICAL	INFINITIVE.SEM_PHYSICAL
	INFINITIVE.SEM_PHYSICAL_OTHER
SENTENCE.DECLARATIVE	SENTENCE.DECLARATIVE
SENTENCE.EXCLAMATIVE	SENTENCE.EXCLAMATIVE
SENTENCE.NONDECLARATIVE	SENTENCE.NONDECLARATIVE
SUBJECT.SEM_ANIMATE_ANIMAL	
	SUBJECT.SEM_ANIMATE_HUMAN
SUBJECT.SEM_INANIMATE_MANMADE	SUBJECT.SEM_INANIMATE_MANMADE

This verification procedure confirms our findings: the best model invests heavily in formal properties such as tense, aspect and mood as well as properties referring to the clause or sentence; together they make up 10 out of 18 properties used. Although tracking these significantly improves prediction accuracy, they are not

typically included in lexical semantic studies while the usual suspects, i.e. semantic properties, seem less reliable predictors for the choice of one near-synonym over another.