DE GRUYTER MOUTON
DOI 10.1515/cog-2013-0008 — Cognitive Linguistics 2013; 24(2): 221–274

Dagmar Divjak and Antti Arppe

Extracting prototypes from exemplars
What can corpus data tell us about concept representation?

Abstract: Over the past four decades, two distinct alternatives have emerged to rule-based models of how linguistic categories are stored and represented as cognitive structures, namely the *prototype* and *exemplar* theories. Although these models were initially thought to be mutually exclusive, shifts from one mechanism to the other have been observed in category learning experiments, bringing the models closer together. In this paper we implement a technique akin to varying abstraction modelling, that assumes intermediate abstraction processes to underlie category representations and categorization decisions; we do so using familiar statistical techniques such as regression and clustering that track frequency distributions in input. With this model we simulate, on the basis of actual usage of Russian TRY verbs and Finnish THINK verbs as observed in corpora, how prototypes for near-synonymous verbs could be formed from concrete exemplars at different levels of abstraction.

In so doing, we take a closer look at the cognitive linguistic flirtation with multiple categorization theories, suggesting three improvements anchored in the fact that cognitive linguistics is a usage-based theory of language. Firstly, we show that language provides support for considering single prototype and full exemplar models as opposite ends along a continuum of abstraction. Secondly, we present a methodology that simulates how prototypes can be obtained from exemplars at more than one level of abstraction in a systematic and verifiable way. And thirdly, we illustrate our claims on the basis of work on verbs, denoting intangible events that are neither stable in nor independent of time and express relational concepts; this implies that verbs are more susceptible to their meanings being influenced by the concepts they relate.

Keywords: categorization; Finnish; exemplar model; frequency distribution; near-synonymy; polytomous logistic regression; prototype model; Russian; usage-based; varying abstraction model

Dagmar Divjak: University of Sheffield. E-mail: d.divjak@sheffield.ac.uk.
Antti Arppe: University of Alberta.

1 Background

A fundamental question in linguistics as well as in cognitive science is how linguistic structures are acquired, how these structures are represented in the human mind, and how they relate to usage as is evident in, e.g. corpora. This question is integrally linked with *categorization*, which is essentially our capacity to classify a continuous stream of external stimuli into a limited number of cognition-internal categories, a.k.a. *concepts*, a capability not restricted to linguistic input alone but most pervasively present in language.

> Categorization is not a matter to be taken lightly. There is nothing more basic than categorization to our thought, perception, action and speech (. . .) And any time we either produce or understand any utterance of any reasonable length, we are employing dozens if not hundreds of categories: categories of speech sounds, of words, of phrases and clauses, as well as conceptual categories. Without the ability to categorize, we could not function at all, either in the physical world or in our social and intellectual lives. An understanding of how we categorize is central to any understanding of how we think and how we function, and therefore central to an understanding of what makes us human (Lakoff 1987: 5–6).

The capacity to classify stimuli into a limited number of categories, i.e. to organize and structure objects in the world around us, is one of the most fundamental abilities in cognitive functioning: categorizing stimuli is one of the cognitive operations (see Bybee 2010: 7–8 for other linguistically relevant general cognitive processes) that make the world more predictable, because many unknown properties of newly encountered stimuli can be induced with sufficient certainty as soon as the stimulus is recognized as a member of a certain category (Estes 1994). The question of how categories are stored in memory and how they are activated in category-related decisions has intrigued psychologists for centuries. By and large, two camps can be distinguished, i.e. rule-based versus similarity-based categorization, and we will introduce each briefly in turn.

1.1 Rule-based categorization

The classical or traditional view, dating from the time of Aristotle, is that categories are mentally represented as sets of individually necessary and jointly sufficient features. According to this view, an exemplar is assigned to a category if it satisfies the category's membership features; following on from the Law of Contradiction and the Law of the Excluded Middle, these features must be binary (Taylor 1989: 23), i.e. they are either present or absent. As a consequence, rule-based categories have clear and predictable as well as seemingly immutable

boundaries that contain all and only those elements that meet the definition. Membership is not a matter of degree.

In linguistics, this logical stance was notoriously represented in phonology, and from there gained ground in syntax and semantics. Following the example of phonology, syntactic and semantic categories were assumed to be represented by means of binary, primitive, universal, abstract and innate features (Taylor 1989: 30). Think, for example, of how a *bachelor* was described by Katz and Postal (1964: 13) as [+human], [+male], [+adult] and [+never married]. All features present in "an unmarried adult human male" are individually necessary and jointly sufficient for being a *bachelor* – women do not qualify, and neither do little boys, husbands or apes. But how about Catholic or Eastern Orthodox priests?

Problems with the traditional approach were noted early on (see Lyons 1977 for a discussion) although most issues with the traditional approach were anticipated by Wittgenstein (1953/2001) in his well-known discussion of how to define the word *Spiel* ('game'). In the case of *game*, there appears to be no set of properties that is shared by all members of the games-category and is exclusive to them such that games can be unambiguously distinguished from non-games. Instead of a "category structured in terms of shared criterial features" (Taylor 1989: 38) we get a complicated network of overlapping and criss-crossing similarities that is learnt on the basis of exemplars, rather than by building up unique and defining feature sets, the applicability of which to any instance can easily be checked. Labov (1973) confirmed Wittgenstein's suspicions experimentally, and further responses to the problem have taken the form of one of two distinct theories, i.e. prototype and exemplar theory, both similarity-based categorization strategies.

1.2 Similarity-based categorization

In the study of lexicalized semantic concepts Eleanor Rosch's (1973 and later work) prototype approach to categorization dominates, i.e. a probabilistic feature approach with instances displaying different degrees of representativity and similarity to a prototype, in its purest version to a single prototype. That prototype representation of a category is generally taken to be a generalization or abstraction of a class of instances falling into the same category. In accordance with the prototype view, a category or concept is thus assumed to be stored as a highly abstract representation, consisting of an aggregate of properties that are characteristic of the concept rather than strictly defining and delineating it. Categorization of a newly encountered entity requires the comparison of that entity with

available prototypes and subsumption of the new entity under the most similar prototype representation.

In contrast, the exemplar view, initially proposed by Hintzman (1986) and Nosofsky (1986), dominates category learning experiments in laboratory settings. On the exemplar view, a category's mental representation encodes the exemplars that compose the category. A category or concept is thus presumed to be represented as detailed memory traces of all the individually encountered exemplars of the concept; in the most extreme version of exemplar theory no abstraction would take place across these stored exemplars. To decide whether an entity is a member of a category, one compares it to the category, i.e. categorization is carried out by analogy to all stored instances.

Although both exemplar and prototype theories come with their own shortcomings (see Murphy 2002; Vanpaemel and Storms 2008: 732–733) they can both account for phenomena that are problematic for the Aristotelian theory, such as the fact that certain members of a category seem to be better members than others. Stimuli will be better exemplars of a category the more closely they resemble the category's prototype, hence all exemplars need not share the same properties. Likewise, stimuli will be better exemplars of a category the more closely related they are to the category's exemplars; if all of the individually encountered exemplars of a category are encoded, they need not share the same properties.

1.3 A usage-based approach to categorization

A considerable literature exists contrasting the prototype and exemplar theories (for an overview, see Nosofsky 1992; for the empirical data, see Murphy 2002; for a theoretical treatment, see Vanpaemel and Storms 2008). Although these two views have been traditionally treated as mutually exclusive and exemplar-models tend to win over prototype models in categorization experiments (for a first comparison see Medin and Schaffer 1978), some stress that exemplar and abstraction models are informationally equivalent (Barsalou 1990), hence cannot be distinguished on the basis of behavioural or reaction time data. Verbeemen et al. (2007: 549), who study lexicalized concepts, would rather see them as opposite ends of a continuum, a conclusion that is very similar to the one reached by Bod (2009); Bod argues with regard to syntactic structures that exemplars and rules are end points of the same distribution. A usage-based view, as adopted in cognitive linguistics, would support this development from mutual exclusion to mutual inclusion and expect prototypes to emerge from repeated exposure to and abstraction over exemplars, just like schemas do.

1.3.1 Reconciling rule-based and similarity-based models of categorization in theory

At first glance, rule-based types of categorization seem absent from cognitive linguistics. Or at least, the word *rule* is conspicuously absent. Instead of rules, cognitive linguists operate with *schemata* (Langacker 1987; compare also Lakoff's 1987 *image schema*), a schema[1] being a "superordinate concept, one which specifies the basic outline common to several, or many, more specific concepts" (Tuggy 2007: 83). In other words, "any concept that abstracts away from differences among similar subcases may be properly called a schema" (Tuggy 2007: 84). Within Cognitive Grammar "any kind of superordinate concept, be it called rule, pattern or template, will be handled by positing a schema or schemata" (Tuggy 2007: 94). Hence, "the highest-level schema [. . .] embodies the maximal generalization that can be extracted as a characterization of the category membership" (Langacker 1987: 380). A schema is thus "an abstract characterization that is fully compatible with all the members of the category it defines" (Langacker 1987: 371). Similarity if not identity of rules and schemas seems implied in the latter, in particular in the specification that "membership is not a matter of degree" (Langacker 1987: 371).

Yet at the same time there is the issue of full versus partial schematicity, depending on whether there is any conflict between the specifications of the sanctioning and target structures. In the case of full schematicity the specifications of both structures are fully compatible while partial schematicity arises when there is some conflict between the specifications of both structures, as is the case in prototype-based categorization (Langacker 1987: 69).

Categorization by prototype plays the more prominent role in another pillar of cognitive linguistics, i.e. Lakoff's (1987) work. Very briefly, the essential conceptual structure for Lakoff's theory is the Idealized Cognitive Model (ICM), the means by which we organize our knowledge. Category structures and prototype effects are by-products of ICM organization. An ICM is a concept or cluster of related concepts that defines our knowledge of a category. There are five types of ICMs (image-schematic, propositional, metaphoric, metonymic and symbolic) and all of them can be roughly defined as mental spaces and models that structure those spaces. The differences lie in what those spaces relate to and the manner in which they are structured. At the center of an ICM are those properties

1 In cognitive psychology, a schema is a cognitive framework or concept that helps organize and interpret information. It structures knowledge, beliefs and expectations, of objects, people and situations, to guide cognitive processes and behaviour. Categorization by schema seems to play only a minor role in the categorization literature, however (Murphy 2002: 47–48).

which most strongly characterize the category. On this approach, members of a category which best fit the definition imposed by the ICM are called prototypes.

It may strike the outsider as odd that cognitive linguists simultaneously subscribe to several different cognitive models of categorization, seen outside of cognitive linguistics as radically opposed (rule versus similarity) or as two different strands within one model (prototype versus exemplar). The traditional cognitive linguistic account of the difference between a prototype and a schema, as cited from Langacker (1987: 371), runs as follows: "a prototype is a typical instance of a category, and other elements are assimilated to the category on the basis of perceived similarity to the prototype". There are thus degrees of membership based on degrees of similarity. "A schema, by contrast, is an abstract characterization that is fully compatible with all the members of the category it defines, so membership is not a matter of degree". Langacker (1987: 380) specifically states that the category prototype has developmental priority and cognitive salience and defines the center of gravity for the category as the primary basis of extension. The highest-level schema, on the other hand, is significant because it embodies the maximal generalization that can be extracted as a characterization of the category membership.

Goldberg (2006) argues that both item-specific exemplar knowledge and generalized or schematic knowledge must be present to account for patterns of learning. Speakers have very specific usage-based knowledge about the constructions they master, including their frequency of occurrence, which makes up item-specific knowledge (cf. Bybee 2010: 18); at the same time, speakers can produce novel utterances, which, according to Goldberg, necessitates generalized or schematic knowledge. It is necessary for these generalizations themselves to be stored, despite the fact that there naturally is partial abstraction in an exemplar-based model, because we do not necessarily record everything about an exemplar and because we may forget the (exact) details of what we did record, even more so as each individual's language may change over time (e.g. Harrington 2006; see also Bybee 2010: 22).

In other words, cognitive linguistics seems to be moving in the direction of integrating similarity-based and rule-based forms of categorization. This is achieved by referring to the fact that elements of a category vary in their cognitive prominence: strongly entrenched and highly salient concepts, also known as prototypes, anchor relations of schematicity (Tuggy 2007: 89). Such a move is facilitated by Langacker's (2010: 132–138) claim that rule- and similarity-based forms of categorization – in the case he discusses schema and exemplar theory – are essentially the same "when stripped of their metaphorical clothing", although they remain distinguishable, and "a model that combines the virtues of the two approaches" is desirable and possible.

Another way of seeing whether different types of categorization are mutually compatible or exclusive is simulating the ways in which categories come into being according to various theories and models. This is the path we will take in this article where we experiment with ways in which prototypes can emerge from exemplars, a process that is illustrated here on the basis of, but is by no means limited to, near-synonymous lexical items; it also applies to research into the existence of local and more general schemata in phonology (Pierrehumbert 2001), morphology (Dąbrowska 2008) and syntax (Verhagen 2005). In so doing, we will address three of the, in our view, major shortcomings of studies on categorization within usage-based frameworks.

1.3.2 Reconciling rule-based and similarity-based models of categorization in practice

Although cognitive linguistics is actively promoted as a usage-based theory, and seems to accept that both details and generalizations are stored redundantly, thus far the way in which prototypes (or schemas, for that matter) are assumed to be extracted from actual usage made up of exemplars has not been modelled computationally.

Furthermore, within cognitive sciences, categorization research on exemplar models has typically used tasks in which novel, artificial categories are learned in the laboratory, limiting the external validity of the results obtained.[2] On the other hand, research concerned with lexicalized semantic concepts referring to structures that exist in the world is hampered by the fact that these structures cannot easily be manipulated. As a consequence, much of the work on prototype-models is correlational in nature (Verbeemen et al. 2007: 538), reducing its explanatory power (in particular the cause-effect relation).

Finally, the bulk of research done on prototype categorization has concentrated on nouns (cf. Pulman 1983). A basic difference between nouns and verbs is that, typically, nouns describe items that are stable in time and therefore independent of that dimension, whereas verbs describe items that are typically neither stable in nor independent of time. In addition, nouns typically denote tangible objects, whereas verbs name intangible events. Furthermore, verbs render relational concepts, which implies that they are more susceptible to their

2 The primary focus has been on the process by which a test stimulus is assigned to one of several contending categories after differences among these categories have been learned. A distinction has generally been made between stimuli that vary in terms of bi-valued and multi-valued dimensions.

meanings being influenced by the concepts they relate. This implies that prototypical situations are partly determined by the elements verbs co-occur with.

It is precisely this contextual element that we aim to exploit in our corpus-based quest for a cognitively realistic and systematic, impartial procedure for extracting verbal prototypes from language use. We do so by statistically modeling large, manually annotated datasets of exemplars and gradually reducing exemplars while abstracting properties. To this end, we implement an idea, similar to that of varying abstraction (Verbeemen et al. 2007), as the basis for categorization. Tests with computational varying abstraction models show that item categorization is improved if an item can be compared to multiple prototypes representing different levels of abstraction: this "balances the opposing pressures of economy and informativeness, providing just enough representational information to describe the category structure in a sufficiently complete way" (Vanpaemel and Storms 2010: 421).

Varying abstraction models come in different guises (Vanpaemel and Storms 2008: 744–745) but all assume that categories are represented by multiple sub-prototypes formed by merging category members together. Our approach with respect to forming clusters is similarity-based (as are RMC [Griffiths et al. 2007], SUSTAIN [Love et al. 2004] and REX in MMC [Rosseel 2002]), and considers any number of interim clusters (as does the VAM [Vanpaemel and Storms 2008, 2010]), although we in practice opt for a smaller number of interim "representations".

An essential difference with the aforementioned varying abstraction models is that our interim representations remain individual exemplars (and implicitly the particular contexts that they contain) rather than aggregates merged over all exemplars within the subclusters. Furthermore, we implement the idea using a multivariate statistical technique designed to deal with categorical variables that have more than two categories, *polytomous logistic regression*, according to the *one-vs-rest* heuristic (see e.g. Arppe 2008, 2009; implementation in R by Arppe 2013). We apply our approach to the study of contextual similarities and differences of two sets of Russian and Finnish near-synonyms expressing TRY and THINK. This can be seen as a rigorous, multifactorial implementation of forms of usage-based analysis such as those presented in Ellis and Ferreira-Junior (2009a, 2009b) and Bybee (2010), going beyond conventional manual scrutiny of concordance lines, frequency counts or comparisons of proportions of occurrence of lexical items in individual contextual slots. Earlier on, Gries (2003b) used a small dataset to demonstrate how a similar multivariate analysis method, *Linear Discriminant Analysis* (LDA), could be used for ordering sentences, representing two constructional alternatives denoting the same meaning, in terms of their prototypicality with respect to the two alternatives (termed *categories*). Gries' approach

effectively merges the concepts of prototype and exemplar by seeing these as manifested primarily in the original sentences (and their constituent properties) in the dataset, and undertakes the prototypicality ordering along a single axis with the two alternatives at the opposite ends. As such Gries (2003b) differs from our approach which distinguishes between the selection of exemplars and the representation of the prototypes (consisting of abstract properties representative of the categories as a whole and not individual sentences), as well as allows for multiple equally exemplary examplars (at the intermediate levels) for a category (which are actual sentences in the dataset).

Likewise, Arppe (2008: 248–252) has explored statistical techniques to select varying numbers of exemplary sentences (incorporating various abstract properties) for a set of synonymous verbs, and has interpreted abstract properties strongly associated with the synonyms as their prototype-like "core semantic characteristics" (Arppe 2008: 160–163), but he does not integrate these separate strands into one unified approach that bridges the prototype and exemplar views as we do in this paper.

The implementation we present confirms that the assumptions underlying the varying abstraction model fit language data well. As such, we go beyond the single prototype models that made their entrance into Cognitive Linguistics in the 80s, beyond the exemplar models that came into fashion with the advent of large corpora in the 90s and beyond hybrid models that aim to combine the best of both worlds by including a mixture parameter weighing the relative contributions of an exemplar and a prototype model (Vanpaemel and Storms 2008: 744); these models have been gaining importance over the past ten years. We thus take up the challenge, put forward by Barsalou (1990: 85), to look for guidance outside the category learning literature and to see what assumptions about representation best serve other cognitive domains, including language, as strategy for furthering research in the domain of category learning.

1.4 Objectives and structure of this study

The objectives of this study are, first, to explore how the prototype and exemplar models of categorization manifest themselves in corpus data, and second, to show that there is support in language for a continuum of abstraction between the two alternative models. Although corpus data do not reflect the characteristics of mental grammars directly, we do consider corpus data a legitimate source of data about mental grammars. Since the results of linguistic cognitive processes, e.g. corpus data, are not independent of, or unrelated to, the linguistic knowledge that is represented in the brain, we may assume with justification that characteristics

observable in language usage reflect characteristics of the mental processes and structures yielding usage, even though we do not know the exact form of these mental representations. Therefore, we feel that corpus data are amenable to our objective of providing evidence for the in cognitive linguistic circles tacitly accepted position that the two major accounts of cognition, typically treated as distinct in the cognitive sciences, are indeed more fruitfully considered as opposite ends of the same distribution. However, psycholinguistic experimentation is needed to validate the exemplar-to-prototype conversion procedure we simulate.

The structure of the paper is as follows. In Section 2 we introduce the data used for the simulation study. In Section 3, we illustrate how, despite the high degree in similarity the near-synonyms in our dataset display, a statistical model manages to predict correctly the choice of one particular verb over one of its near-synonyms in more than half of all cases. It is contextual information, i.e. structural and lexical information found within sentence boundaries, which allows these models to do so. For Russian, morphological Tense-Aspect-Mood markers as well as the semantics of the subject and infinitive are crucial; for Finnish, the morphological makeup of the verb and the verb-chain it is part of, in combination with the semantic classification of the syntactic arguments linked to these verbs, prove essential. In Section 4 we provide corpus-based evidence supporting the hypothesis that every encounter with a verb contributes to a "cloud" of exemplars from which salient properties are extracted. These properties become part of the lexeme-specific information for each verb – whether it is called ICM or prototypical usage – and are inferred even when those salient contextual properties are not available. Section 5 summarizes our findings while outlining avenues for further research. The method outlined in this paper is available upon request as a vignette with stage-by-stage instructions and accompanying R functions (Arppe 2013).

2 Data collection

In the spirit of recent developments within corpus-based approaches to cognitive linguistics (Grondelaers et al. 2002; Gries 2003a; Bresnan et al. 2007), large-scale corpus-based studies (Arppe 2008; Divjak 2010) have explored the phenomenon of near-synonymy from different angles. A whole battery of statistical techniques, ranging from exploratory techniques to full-blown multivariate predictive models, has been called into service to explore the similarity of near-synonymous items and study the contextual similarities and differences of a set of six near-synonymous verbs denoting TRY in Russian and four THINK verbs in Finnish.

This interest marks a change with respect to the second half of the 20th century, when near-synonymy did not receive a great deal of attention. From the point of view of categorization, however, near-synonymy would seem to be the norm rather than the exception: similar experiences that can nevertheless be distinguished may well deserve different labels. As an introductory illustration, let us look at an example for each of these near-synonym sets that contains, within the same sentence or paragraph more than one of the near-synonyms for each of the languages. Example (1) contains two out of four Finnish THINK-verbs, *miettiä* and *ajatella*, while example (2) contains three out of six Russian TRY-verbs, i.e. *silit'sja*, *starat'sja* and *probovat'*.

(1) En halua esittää mielipiteitä **miettimättä** tarkasti, mitä oikeastaan **ajattelen**. I do not want to present opinions **without thinking** carefully, what I actually **think** [about them]. [sfnet.keskustelu.ihmissuhteet]

(2) Но Сирота все еще **силился** что-то сказать, и снова невозможно было понять ни слова из того, что он говорил. Малинин наконец не выдержал и прекратил эту обоюдную муку:
Ты **не старайся**, Сирота, все равно я не понимаю: у тебя рот разбитый . . . Звук и только, а голоса нет. В госпитале полежишь – восстановится, а сейчас **не пробуй**, не мучь себя (. . .). [K. Simonov. Živye I Mertvye]
But Sirota was still **trying** [silit'sja] to say something, and again it was impossible to understand a word of what he was saying. Finally, Malinin could not take it any longer and put an end to this mutual torture:
"**Don't you try** [starat'sja], Sirota, I can't understand you anyway: your mouth got smashed . . . . There is only sound, no voice. You'll be in hospital for a while – it will heal, but for now **don't try** [probovat'], don't torture yourself." (. . .)

In a semiotic system characterized by limited lexical resources, near-sameness of meaning is often considered aberrant: instead of economizing the expressive potential of the language by making a single lexical item express multiple meanings, near-synonymy apparently decreases the language's expressive efficiency by allowing several lexical items to convey (roughly) the same meaning. But how do speakers decide on the cut-off level for assigning experiences a different label naming "roughly" the same experience? How do speakers come to realize that words express "(roughly) the same meaning" and how do they find the differences? And once these labels are in place, what does expressing "(roughly) the same meaning", i.e. near-synonymy, really mean?

The datasets collected and annotated for TRY verbs in Russian (Divjak and Gries 2006; Divjak 2010) and THINK verbs in Finnish (Arppe 2008) form the basis for exploring these issues; the data sets will be briefly introduced in the following sections, with full specifications provided in the two monographs referred to.

2.1 Data[3]

Contextual data on the six most frequent Russian verbs that express TRY when combined with an infinitive, i.e. *probovat', pytat'sja, starat'sja, silit'sja, norovit', poryvat'sja,* were extracted from corpora. The main source of data is the ten-million-word section (10,750,757) of the Amsterdam Corpus (AC) that contains literary works from different genres, originally written by approximately 75 authors in Russian between 1950 and 2000. This dataset was later supplemented where necessary with data from the Russian National Corpus (RNC), when it became available, to result in approximately 250 extractions per verb to the extent possible. Data was extracted from the RNC from literary works from the same period as contained in the AC. The exact numbers of examples that were annotated per verb studied are given in Table 1. Depending on the frequency of the verb, between 119 and 260 examples were annotated. In all, there were 1,351 occurrences of this syntactically homogeneous category, that is, a category in which all TRY verbs share the same argument structure, i.e. they all require an infinitive. Sociolinguistically speaking, the category is less homogeneous, with *silit'sja* and *norovit'* belonging to spoken language (Ožegov and Švedova 1999).

Table 1: Corpus examples used per TRY verb

Verb	N (AC/RNC)
probovat'	246 / –
pytat'sja	247 / –
starat'sja	248 / –
silit'sja	57 / 185
norovit'	112 / 148
poryvat'sja	31 / 88

3 All data collected could be subsumed under the header "contextual data" if we take into account that (morphological and semantic) information embedded in the word itself, i.e. the "internal context" in one language may have a parallel in other languages in the conventional "external context" (cf. Arppe 2002).

Table 2: Corpus examples used per THINK verb

Verb	N (HS/SFNET)
ajatella	1492 (570 / 922)
miettiä	812 (335 / 457)
pohtia	713 (556 / 157)
harkita	387 (269 / 118)

For Finnish, the four most frequent synonyms meaning 'think, reflect, ponder, consider', i.e. *ajatella, miettiä, pohtia, harkita*, were extracted from two months of newspaper text from the 1990s (Helsingin Sanomat 1995) and six months of Internet newsgroup discussion from the early 2000s (SFNET 2002–2003), namely regarding (personal) relationships (sfnet.keskustelu.ihmissuhteet) and politics (sfnet.keskustelu.politiikka). The newspaper corpus consisted of 3,304,512 words of body text (i.e. excluding headers and captions as well as punctuation tokens), and included 1,750 examples of the studied THINK verbs. The Internet corpus comprised 1,174,693 words of body text (excluding quotes from previous postings), yielding 1,654 instances of the studied THINK verbs. In terms of distinct identifiable authors, the newspaper sub-corpus was the product of just over 500 journalists and other contributors, while the Internet sub-corpus involved well over 1000 discussants.

In all, there were 3,404 occurrences of this syntactically non-homogenous category (i.e. not all verbs share exactly the same argument structure), with frequencies ranging from 1,492 for the most common one *ajatella* to 387 for the rarer *harkita*.

2.2 Annotation

For Russian, the 1,351 examples were tagged using the annotation scheme proposed in Divjak (2004). This scheme captures virtually all information provided at the clause or sentence level by tagging morphological properties of the finite verb and the infinitive, syntactic properties of the sentences and semantic properties of the infinitive as well as optional elements. There were a total of 14 multiple-category variables amounting to 87 distinct variable categories or contextual properties[4]. For details we refer to Table 3 in Section 3.1

4 In this paper we use the more general term "(contextual) property" as an umbrella for the terms we have used in earlier analyses of these verbs, i.e. "ID tag" in work on Russian verbs of *trying* and "feature" in work on Finnish verbs of *thinking*. The terms "variable" and "variable

The tagging scheme was built up bottom-up, and results from an incremental experience- and distribution-based grammatical- and lexical-conceptual analysis. Although we do assume that native speakers of Russian have a sense of past, present and future, of aspectual oppositions and other, semantic, properties tagged for, we do not, however, claim that native speakers would operate with these exact same categories, let alone label these categories the way that is done here. This conclusion is supported by our finding that quite similar levels of model fit and prediction accuracy can be achieved by selecting clearly divergent sets of properties in a model, even when undertaking this selection at random (see Section 3.2).

For Finnish, the 3,404 examples were first morphologically and syntactically analyzed using an implementation of Functional-Dependency Grammar (Tapanainen and Järvinen 1997; Järvinen and Tapanainen 1997), namely the FI-FDG parser (Connexor 2007). Next, all the instances of the studied verbs, together with all their relevant associated context (not limited merely to obligatory syntactic arguments), were manually checked, corrected and supplemented with semantic subclassifications. A complete overview of the properties tagged for in the analysis of the corpus data is presented in Table 4 in Section 3.1. In all, 477 contextual properties or property combinations were retained as they were sufficiently frequent.[5]

Although the two analysis schemes have different starting points (i.e. an argument structurally homogeneous category for Russian versus an argument structurally varied category for Finnish) and, as a result, operate with a different set of analytical categories, they are nevertheless similar in trying to grasp the immediate context of a verb in its entirety. Moreover, using two distinct schemes dealing with distinct near-synonym sets in distinct languages is a test of the overall robustness of the statistical modeling and analysis, provided we are able to produce effectively similar results.

Before proceeding to the model, a caveat is in order. The use of an annotation scheme may lead readers to conclude that the analysis does not really begin from exemplars, but rather with properties abstracted by the analysts from the examples. While it is true that the polytomous logistic regression model works with abstractions (representing sets of individual words or phrases), the abstractions

category" are used for referring to the annotation of the dataset, with "variable" equalling "(contextual) property" or "feature" and "variable category" being equal to ID tag category. The term "parameter (value)" is used in the context of statistical analysis, with "parameter" referring back to "(contextual) property".

5 Established as $n \geq 24$ according to the so-called Cochran conditions (Cochran 1952, 1954) at the univariate analysis stage.

themselves are the results of a bottom-up process that starts with the linguistic analysis of each corpus extraction with its unique combination of words and phrases. The chosen labels for the Russian dataset employ superordinate terms for the instances encountered, e.g. *animate* to summarize *cats, dogs* and other non-human animals, while the labelling for Finnish reflects the reality that the variables are combinations of many distinct contextual elements sharing essential semantic properties. Moreover, it should be noted that other multivariate methods such as memory-based learning or naïve discriminative learning, deemed cognitively more plausible, are often applied using similarly abstracted variables instead of starting with the raw, unclassified instances of words or phrases at the individual sentence level.

In the next section, we briefly illustrate how the model was built and how it performs. This provides the necessary basis for trusting the relation it suggests between exemplars and prototypes in Section 4.

3 Statistical models and the prediction of synonym choice

Three related questions are relevant to any quantitative, usage-based framework of linguistic research. First, what is the nature of the relationship between naturally produced language and the posited underlying language system that governs such usage? Can the use and choice among lexical and structural alternatives in language be accounted for by referring to underlying explanatory factors, within the framework of some linguistic theory, be it based on categorical generative rules or probabilistic regularities that are assumed to represent language as a system? Second, how can this relation between structure and use be modeled using statistical methods? The complexity of linguistic data necessitates the use of multiple linguistic variables from a wide range of categories, instead of only one or two, and hence motivates the use of multivariate statistical methods. Since modeling a linguistic phenomenon thus becomes a matter of degree, this leads to the third question of how much of actual, real usage can ultimately be modeled accurately? Despite the high degree in similarity that most near-synonyms display, statistical models manage to predict the choice of one particular verb over one of its near-synonyms correctly in more than half of all cases on the basis of explicit intralinguistic information alone[6], and in fact better, if we generalize and

6 But see Inkpen and Hirst (2006: 25–27) and Inkpen (2004: 111–112) for close to perfect prediction results when including pragmatic and extralinguistic variables such as denotational

look at overall proportions of occurrences in particular distinct contexts rather than at individual choices.

We provide an answer to these three questions by fitting a (polytomous) logistic regression model to the annotated near-synonym data (presented in Section 2). Logistic regression looks at outcomes as proportions among all observations with the same context rather than as individual *either-or* dichotomies of occurrence vs. non-occurrence. In other words, logistic regression estimates probabilities of occurrence given a particular context. The variable parameters (typically designated in regression equations as Greek betas, *β*'s) it estimates can be interpreted naturally as odds (i.e. *exp[β]*) (Harrell 2001). Our research question thus translates into the question of how much the presence of a contextual property increases (or decreases) the *chances* of obtaining a particular outcome (i.e. a particular synonym), with all the other explanatory variables being equal.

Logistic regression, and binary logistic regression in particular, have been applied to a variety of linguistic phenomena starting with variable-rule (VARBRUL) modeling in variationist linguistics (e.g. Sankoff 1978[7]; Grondelaers et al. 2002; Bresnan et al. 2007). For a variety of reasons we use polytomous logistic regression with the one-vs-rest heuristic (Rifkin and Klautau 2004; Arppe 2008). As opposed to standard multinomial logistic regression, the one-vs-rest heuristic distinguishes each member of the category from the rest, instead of contrasting it against an individual baseline category. This is desirable because it obviates the need to assume that any single verb of the two synonym sets would be the most prototypical case against which all the others are contrasted. Instead, we are interested in which contextual features distinguish all verbs from the rest in the synonym sets, including a possible "prototype" verb. Furthermore, the one-vs-rest heuristic directly provides lexeme-specific odds with respect to selected variables, representing linguistic properties. As a simple selection rule, we then pick the verb receiving the highest probability in a given context.[8]

---

micro-distinctions as well as expressive ones concerning the speaker's intention to convey some attitude, in addition to the sought-after style. Note, however, that these properties are collected from dictionaries containing synonym differences rather than from free text in which the synonyms occur, though they are assumed to be inferable from corpus data (Inkpen and Hirst 2006: 35).

7 Sankoff (1978) does not yet use the term *logistic regression* even though the mathematical specification is clearly a logistic regression model.

8 The highest estimated probability is not necessarily always close to $P = 1.0$ or even $P > 0.5$ but can range from slightly over $1/n$ (n indicating the overall number of outcomes) to 1.0, i.e. from slightly over ⅙ to almost 1.0 for the Russian TRY lexemes and from slightly over ¼ to almost 1.0 for the Finnish THINK lexemes (cf. Arppe 2008: 129–130; Hosmer and Lemeshow 2000: 156–60).

Before proceeding to the analysis, a caveat relating to cognitive reality needs to be expressed. We aim to model produced language by means of a statistical heuristic, logistic regression analysis. This technique was chosen because, different from most other machine learning methods, it makes it possible to find out what is happening "internally", i.e. it is possible for the researcher to extract linguistically meaningful weights on the contextual explanatory variables the model is operating on (cf. Harrell 2001). Being able to take "a look under the hood" at every stage in the abstraction process is a *sine qua non* in constructing a varying abstraction model. The heuristic by which our polytomous model is constructed and the constituent binary logistic regression models and mathematical algorithms by which the models are optimized to fit the data were not primarily designed to mimic cognitive or neurological behaviour, however. This is the case for the vast majority of statistical analysis and learning techniques currently used in language research, including the much acclaimed connectionist models (see Hawkins [2004: 24, 37] on the neurological implausibility of such models; cf. also Baayen 2007).[9] This being said, our resulting models fit descriptions that linguists feel are appropriate for the data and the underlying mechanics of regression analysis has indirect cognitive grounding in the fact that human beings are able to detect statistical regularities in input (Saffran et al. 1996, to name but one), although people seem to need much less data than a regression model requires (see Carey and Bartlett 1978 for word learning) and put fewer constraints on variable selection.

A model created with (polytomous) logistic regression provides probability estimates for the (proportional) occurrence of an outcome, in this case a verb within a synonym set, given the contextual occurrence of a combination of linguistic properties incorporated in the model, using the odds assigned to the properties in question as a result of fitting the model to the data. We will illustrate this process in Section 3, before proceeding to modeling the relation between exemplars and prototypes in Section 4. First, we show how the model was fitted

9 The one exception we are aware of is naïve discriminative learning (NDL) that can be applied to symbolic abstractions of linguistic cues for which relative weights can be estimated (Arppe and Baayen 2011; Baayen 2011; Baayen et al. 2011) and which has become available only recently (R package ndl, Arppe et al. 2012). However, in a comparison of various machine-learning methods (including polytomous logistic regression with both fixed and random effects), Arppe and Baayen (2011) found that polytomous logistic regression and naive discriminative learning appeared to be closest in terms of the predictions that the models make. It would seem that these two models are effectively achieving the same results, although NDL does this in a cognitively more plausible way and we intend to use it in future work.

to the data. Next, we take a look at the performance the model for both languages, then move on to illustrating what type of predictions the model makes, and how it arrives at its decisions. These are necessary steps in assessing the confidence we should have in a model that simulates how prototypes could be abstracted from exemplars.

3.1 Fitting a model

The original annotated dataset of the Russian TRY verbs (Divjak and Gries 2006; Divjak 2004, 2010) contained 14 multinomial variables corresponding to 87 binary true/false variable categories. These had to be pruned, since the number of variable categories allowed in (polytomous) logistic regression is maximally 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 six 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.[10] 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 with all six 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* (Theil 1970) value (a measure of nominal category association) larger than 0.5 (i.e. one variable reduces more than ½ of the uncertainty concerning the other).[11] Altogether 18 variable categories were retained (11 semantic versus 7 structural) and are listed in Table 3.

10 Among the alternative models that were fit, one model was based on 15 variables with properties occurring altogether at least 45 times, individually with at least 3 TRY lexemes, and 20 times with at least 2 TRY lexemes. This allows for, and indeed leads to, slightly clearer verb-specific preferences, the accuracy rate being 51.07%.

11 We are aware of the general tendency to use statistical metrics to reduce the number of variables (step-wise variable selection, Variable Inflation Factor [VIF], etc.). However, we have chosen not to do so since such quantities are all too often used to "outsource" model specification to the software instead of relying on the domain knowledge of the researcher(s) (Harrell 2001). Moreover, O'Brien (2007) notes that the suggested remedial actions to be undertaken/required when whatever metric of multicollinearity exceeds some (in effect arbitrary) threshold are questionable.

Table 3: The 18 contextual variables retained for modeling the Russian data

Variable class	Properties	Number of properties included in the model	Number of properties included in the annotated dataset
Clause type	Main (vs. subordinate)	1 [main clause]	2 [main vs. dependent]
Sentence type	Declarative (vs. other rarer types)	1 [declarative]	4 [declarative, exclamative, imperative, interrogative]
Finite verb: morphological properties	Aspect, tense, mood	4 [perfective verb, indicative and gerund mood, past tense]	12 [present, past, future; infinitive, indicative, subjunctive, imperative, participle, gerund; imperfective vs. perfective]
Infinitive verb: morphological properties	Aspect	1 [imperfective verb]	2 [imperfective vs. perfective]
Infinitive verb: degree of control	High (vs. medium and low)	1 [high control]	3 [High vs. medium vs. no controllability]
Infinitive verb: semantic characterization	Communication, exchange, motion, metaphorical motion, etc.	9 [communication, exchange, physical (involving self), physical (involving other), motion (involving self), motion (involving other), metaphorical motion, metaphorical physical exchange, metaphorical physical involving other]	14 [physical actions, perception, communication, intellectual activities, emotions, wishes/ desires etc.]
Syntactic subject: semantic characterization	Animate human vs. rarer other types	1 [animate human]	9 [concrete vs. abstract, animate (human, animal) vs. inanimate (event, phenomenon of nature, body part, organization/ institution, speech/text) etc.]
	Adverbs & particles	0	[time, location, duration, repetition, intensity etc.; exhortation, permission, restriction etc.]
	Negation	0	[present vs. absent, attached to finite verb or infinitive]

For Finnish, a similar pruning procedure was undertaken, since a maximum of approximately 40 variable categories was allowed, based on the fact that the least frequent verb, *harkita*, occurred 387 times. The remaining variable categories are presented in Table 4. Similar to the selection process for Russian, a minimum overall frequency was required, in this case set at $n \geq 24$. Pair-wise associations of individual properties were likewise carefully evaluated, but due to the heterogeneity of the argument structure of the Finnish THINK verbs, occurrence with all four verbs was not required. In general, only verb-chain general properties and their most general semantic classifications were included in the model, excluding node-specific ones. Of the most frequent syntactic arguments (and/or functional dependents), i.e. PATIENT, AGENT, MANNER, and TIME-POSITION, semantic and structural subtypes were incorporated in the model at the highest level of granularity. Rarer cases, i.e. META-comments (clause-adverbials), LOCATION, CO-ORDINATED VERBS, DURATION, FREQUENCY, QUANTITY, SOURCE, and GOAL, were incorporated simply and uniquely as a syntactic argument type, even if semantic or structural subtypes were available for them. Extra-linguistic properties were excluded. Altogether 46 variable categories were retained for the Finnish model.

Note that our models do not include interactions. Since the number of possible dummy variables is double or more than what the data allows (as discussed in Section 3.1), we have opted to include simple predictors only in the models; favouring a parsimonious yet adequate model over a more complex one is standard practice in statistical analyses. Moreover, a tentative trial of all possible two-way interactions in Arppe and Baayen (2011) produced only a minimal gain but an extremely overfitted, in fact uninterpretable, model.

3.2 Model performance

Table 5 summarizes model performance for Russian and Finnish, as calculated on the basis of the summary data presented in Tables 6 and 7, in two measures.[12]

12 Two further measures can be reported. The measure for the proportionate reduction of prediction error, designated as $\lambda_{prediction}$ (Menard 1995: 28–30), tells us that the models perform 41% (for Russian) and 37% (for Finnish) better by using the selected set of linguistic explanatory variables than what would be achieved by systematically selecting the most frequent verb in the data (which would yield at most 248/1351 or 18.3% correct predictions with the TRY verbs and 1492/3404 or 43.8% correct predictions for the THINK verbs). The measure for proportionate reduction of classification error, designated as $\tau_{classification}$ (Menard 1995: 28–30), informs us that the models reproduce, in the long run, the actually occurring outcome proportions evident in the data better than the baseline case of homogeneous proportionate distribution among outcomes by, respectively, 42% for the Russian case and 49% for the Finnish case.

Table 4: The 46 contextual variables retained for modeling the Finnish data

Variable class	Properties	Number of properties included in the model	Number of properties included in the annotated dataset
Verb-chain: general morphological features	Polarity, Mood, Voice, Person, Number, Explict vs. Implicit subject, Clause-equivalent	10 [negation, indicative, conditional, passive, first, second, third, plural, implicit-subject, clause-equivalent]	24 [affirmation, imperative, active, singular, 1st + singular, 2nd + singular, 3rd + singular, 1st + plural, . . .]
Verb-chain: semantic characterizations	Different types of modality	6 [Possibility, Necessity, External (cause), Volition, Temporal, Accidental]	22 [. . . vs. Propossibility, Impossibility, Obligation, Nonnecessity, Futility, Boldness, Energy, Permission, Prohibition, Ability, Tentative, Start, . . .]
Syntactic argument types alone	Source, Goal, etc.	10 [Source, Goal, Quantity, Location, Duration, Frequency, Meta-Comment, Reason/Purpose, Condition, Co-ordinated-Verb	22 [. . . vs. Agent, Patient, Manner, Time]
Syntactic argument types + semantic/ structural subtype classifications	Agent + Individual, Agent + Group, etc.	20 [Agent + Individual, Agent + Group, Patient + Individual/ Group, Patient + Abstraction, Patient + Activity, Patient + Event, Patient + Communication, Patient + IndirectQuestion, Patient + DirectQuote, Patient + Infinitive, Patient + Participle, Patient + *että*, Manner + Generic, Manner + Frame, Manner + Positive, Manner + Negative, Manner + Agreement, Manner + Joint, Time + Definite, Time + Indefinite]	156 [. . . vs. Quantity + Much, . . . , Location + Event, . . . , Duration + Long, . . . , Frequency + Often, . . . , Co-ordinated-Verb + Mental, . . .

Table 5: Model performance for Russian and Finnish

Model	R_L^2	Accuracy (%)
Russian (one-vs-rest)	0.31	51.7 (699/1351)
Finnish (one-vs-rest)	0.31	64.60 (2199/3404)

R_L^2 is an indicator of how well (or how poorly) a logistic regression model, specifically its probability estimates, fits with the actually observed occurrences in the original data (Hosmer and Lemeshow 2000: 165–166). It is reassuring to see that the model fits the actual occurrences in the original data quite well and does so equally well in both Russian and Finnish.

The accuracy value, in turn, tells us how many correct classifications there were according to the selection rule. In the Russian case, 51.7% of all instances were correctly classified.[13] Before dismissing this figure as "hardly more than half", firstly be reminded that we are predicting a 6-way choice between near-synonyms. In a 4-way choice between semantically related verbs that would not qualify as near-synonyms (i.e. *impose*, *believe*, *request* and *correlate*), the average non-English US college applicant got 52.7% correct (Landauer and Dumais 1997). This result shows that the statistical model is performing well, especially when considering that the task is more demanding than the one the L2 learners were presented with in two respects: the verbs are close synonyms and there are six of them to choose between. Secondly, we should remember that classification as a task is categorical in nature and masks the underlying probabilities, especially in a polytomous setting with more than two alternatives (recall the discussion of the smallest maximum Probability from Section 3, Footnote 8). Thirdly, from a linguistic perspective, since we are dealing with synonymous sets of lexemes, we may expect relatively similar underlying probabilities instead of significant dispersion in their values, since, in principle, any of the alternative synonyms can be used in most, if not all, of the studied contexts, albeit with slightly distinct semantic associations. For the Finnish THINK verbs, all the aforementioned applies as well, with the corresponding minimum approximately equal probability with four possible alternative outcomes being naturally $P > 1.0/4 \sim 0.25$.

Given that both the fit of the models and their prediction efficiency were evaluated using the entire Finnish corpus ($n = 3404$) and the entire Russian corpus

13 Compare these results with a baseline-category model with 26 variables (in which only complementary variables were excluded) which reaches an accuracy of no more than 53% (719/1351 correct choices) (Divjak 2010).

($n = 1351$) as testing data, i.e. the same data which had been used to train the models, the performance results could be considered somewhat optimistic. Nevertheless, validating the Finnish model using 1000-fold simple bootstrap resampling yields only slightly lower performance figures, being mean $R_L^2 = 0.287$ with 95% Confidence Interval CI = (0.264, 0.300), and overall accuracy = 63.8% with 95% CI = (63.1%, 64.5%). The same holds for the Russian model, with a 1000-fold simple bootstrap yielding a mean $R_L^2 = 0.272$ with 95% Confidence Interval CI = (0.237, 0.291), and overall accuracy = 50.26% with 95% CI = (48.93%, 51.44%).[14]

3.3 The classification table

Predictive success can also be assessed by looking at the classification table, showing correct and incorrect classifications of the polytomous dependent. The resulting raw classification table for the six Russian TRY-verbs is presented below (Table 6). This table, read horizontally, contains data on how often which prediction was made according to the selected 18-property model, given that all the available examples were used. The first column of Table 6 contains the six verbs studied, while the first row features the six outcomes that can be predicted. The boldfaced numbers indicate how many times the correct verb was predicted, i.e. how often a certain prediction was made and borne out by the data, e.g. for 143 out of all 250 examples featuring *norovit'*, *norovit'* was indeed predicted based on the information contained in the context. For 32 examples *poryvat'sja* was incorrectly suggested, in 4 cases *probovat'*, in 36 cases *pytat'sja*, for 17 instances *silit'sja*, and in 18 instances *starat'sja*. Information for the remaining five verbs should be read in the same way.

14 For Finnish, a 100-fold random selection of 46 variables from an extended 62-variable set was run. The mean accuracy for these 100 random models was 60.75%, ranging from 54.61% to 64.42%. While the best Finnish random model had as many as 31 (67.4%) variables in common with the model used in this paper, the worst random model still shared 28 variables (60.9%) with the best model. Comparing the best and worst performing random variable sets to each other showed that they had 32 variables in common (69.6%).

For Russian, a 1000-fold random selection of 18 variables from the original full 26-variable set was run. The mean accuracy for these 1000 random models was 45.95%, ranging from 26.87% to 51.59%, thus having a much broader range compared with the corresponding Finnish results, with similar maximum accuracy values but also clearly lower accuracy rates. 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.

Table 6: The classification table for Russian

Original/ Predicted	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**

In all cases, the original verb was indeed predicted most frequently. This resulted in a correct raw classification rate of 51.74%, i.e. in 699 sentences the correct verb was predicted. This number is much higher than what would be expected merely by chance. Given that we are dealing with a 6-way choice and 1351 choices to be made, one could in principle assume that a rate of 16.6% correctly classified instances could be achieved by chance (or slightly higher at 18.5%, if the most frequent TRY verb *probovat'* is always selected).

A similar exercise can be undertaken for Finnish, shown in Table 7. Table 7 contains distributions of predicted against original lexemes using the selected model on the entire data set ($n = 3404$). Here too, for each original lexeme the most frequently predicted outcome is always the lexeme itself. Likewise, for each predicted lexeme, overall, the lexeme itself accounts for the largest proportion of original occurrences.

Table 7: The classification table for Finnish

Original/Predicted	ajatella	miettiä	pohtia	harkita	∑(original)
ajatella	**1275**	97	62	58	1492
miettiä	235	**377**	143	57	812
pohtia	145	144	**365**	59	713
harkita	103	48	54	**182**	387
∑(predicted)	1758	666	624	356	**3404**

3.4 The results in some detail

The prediction tables, Tables 6 and 7, show that the properties included in the analysis do a decent job capturing the differences between the verbs, in both Russian and Finnish.[15] They also reveal information about the relations between these verbs, indicating which ones are overall, usage-wise, more or less similar to each other.

We can go into more detail here by going down to the property level to compare verbs. Consider the following examples illustrating three highly similar Russian verbs from one of the clusters identified in Divjak (2004) and Divjak and Gries (2006), i.e. *pytat'sja*, *starat'sja* and *probovat'*. In (3), the probability of finding *probovat'* given the context, detailed between curly brackets immediately following the example, is 0.63.

(3) Можно. Я вас даже покатаю на катере. Кстати, у вас изумительный голос. Вы никогда не пробовали/**probovali** петь? [Evgenij Kukarkin. Princ. #315]
By the way, you have a wonderful voice. Have you ever tried [probovat'] singing?

{CLAUSE.MAIN, FINITE.MOOD_INDICATIVE, FINITE.TENSE_PAST, INFINITIVE.ASPECT_IMPERFECTIVE, INFINITIVE.CONTROL_HIGH, INFINITIVE.SEM_PHYSICAL, SUBJECT.SEM_ANIMATE_HUMAN}

This probability estimate of 0.63 aggregates multiple predictor odds-ratios. It is calculated on the basis of the property-wise odds listed in Table 8 below, which have been fitted to best match the data taken as a whole (instead of individual instances), using the formulas presented in (4–5) as follows:

15 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. Very different stored property combinations making up the core of the prototype would result in prototypes being different for every person and would make it irrelevant what learners track, as long as they track something. On the other hand, largely overlapping property combinations making up the core of a prototype, which the results of these random variable selection results point to, would make it possible for speakers to draw similar interpretations regarding prototypes 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 the abstract prototypes derived from such sets of properties, is the topic of ongoing research (Arppe et al. In preparation).

(4) P(*probovat'*|Context)/P(¬*probovat'*|Context)
= 1:22 ~ Intercept
· 3.4:1 ~ CLAUSE.MAIN
· 1:2.8 ~ FINITE.MOOD_INDICATIVE
· 1:1 ~ FINITE.TENSE_PAST
· 6.1:1 ~ INFINITIVE.ASPECT_IMPERFECTIVE
· 1:1.2 ~ INFINITIVE.CONTROL_HIGH
· 3.9:1 ~ INFINITIVE.SEM_PHYSICAL
· 1.5:1 ~ SUBJECT.SEM_ANIMATE_HUMAN
= 1/22 · 3.4/1 · 1/2.8 · 1/1 · 6.1/1 · 1/1.2 · 3.9/1 · 1.5/1
= 1.64/1

(5) P(*probovat'*|Context)
= 1.64/(1 + 1.64)
≈ 0.62 → 0.63 (due to the adjustment so that $\sum_{verb}$P(*Verb*|Context) = 1.0)

The probabilities for the other five TRY verbs for exactly the same context can be calculated in a similar manner from the corresponding lexeme-specific property-wise odds in Table 10. The resulting probabilities are the following:

P(*starat'sja*|*Context*) = 0.14
P(*norovit'*|*Context*) = 0.11
P(*pytat'sja*|*Context*) = 0.06
P(*poryvat'sja*|*Context*) = 0.05
P(*silit'sja*|*Context*) = 0.01

It is clear that the estimated probability of encountering *probovat'* (0.63) is much higher than that of any of its five contenders: *starat'sja* follows with *P* = 0.14 and *norovit'* with *P* = 0.11 while the likelihood of encountering any remaining three verbs in this context remains below *P* = 0.06. In other words, we can predict not only which verb is most favoured given a particular context, but also which verb is second, third, fourth, fifth and sixth choice given the context in question – the method not only zooms in on the winner but also provides information on the runners-up. Furthermore, we can explain why a particular verb was preferred over the others in a particular context by looking at the odds per property per verb, something we will do below.

For sentence (6) incorporating the contextual properties listed in (7), the probability of finding *starat'sja* is 0.78, as (8) shows.

(6) Страшно. Ну и что . . . с ним? _ спросил, стараясь/**starajas'** не выдавать голосом своего волнения, некий невидимый из-за спин, торсов [. . .] [E. Popov. Samolet na Kel'n. #1265]
Terrible. So what's the matter with him? he asked, trying [starat'sja] not to let the agitation show in his voice, while remaining invisible behind all the backs . . .

(7) {FINITE.MOOD_GERUND, INFINITIVE.ASPECT_IMPERFECTIVE, INFINITIVE.CONTROL_HIGH, INFINITIVE.SEM_METAPHORICAL_PHYSICAL_EXCHANGE, SENTENCE.DECLARATIVE, SUBJECT.SEM_ANIMATE_HUMAN}

(8) norovit'	poryvat'sja	probovat'	pytat'sja	silit'sja	**starat'sja**
0.06	0.08	0.01	0.05	0.02	**0.78**

And, finally, for sentence (9) incorporating the contextual properties listed in (10), the probability of finding *pytat'sja* is 0.60 (see 11).

(9) (. . .) раздраженно взбил траву, еще раз недобро покосился на Нюшку, сплюнул и, вспомнив, что рядом чужой человек, попытался/**popytalsja** усмехнуться. [V. Solouchin. (Source text not listed). #1034]

"(. . .) irritated he kicked the grass, again he cast a sidelong unfriendly look at Njuška, spat and, having remembered that a stranger was next to him, he tried [pytat'sja] to smile."

(10) {FINITE.ASPECT_PERFECTIVE, FINITE.MOOD_INDICATIVE, FINITE.TENSE_PAST, INFINITIVE.CONTROL_HIGH, SENTENCE.DECLARATIVE, SUBJECT.SEM_ANIMATE_HUMAN}

(11) norovit'	poryvat'sja	probovat'	**pytat'sja**	silit'sja	starat'sja
0	0	0.20	**0.60**	0	0.21

The odds used in the calculations were taken from Table 8 below, which summarizes the verb specific odds per property for all six Russian verbs. Bold-faced odds (>1) are significantly positive odds, i.e. in favor of a lexeme, odds in round brackets are insignificant, i.e. neutral, whereas fractioned odds (<1) indicate odds significantly against a lexeme. Take for example the property CLAUSE.MAIN, exemplified in sentence (3). This property has significant positive odds in favor of *probovat'* (3.4:1), neutral ones for *silit'sja*, *starat'sja*, *norovit'* and *poryvat'sja*, and significant odds against *pytat'sja* (1:1.6). Moreover, the comparatively high odds

of 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.

For the Finnish THINK verbs the summary of verb-specific property-wise odds is presented in Table 9 below. The number of significant odds per verb appears to be associated with the overall frequency of the lexeme, as for the most frequent *ajatella* 32 properties overall exhibit significant odds either in favour of or against it, while the respective figures for the rarer lexemes are 22 for *miettiä*, 20 for *pohtia*, and 13 for *harkita*. More specifically, among the significant odds for each verb, 15 are in favour of and 17 against the occurrence of *ajatella*, whereas the corresponding figures are 14 vs. 8 for *miettiä*, 12 vs. 8 for *pohtia*, and 6 vs. 7 for *harkita*. Thus, the balance of properties in favour of or against a verb varies, with *miettiä* and *pohtia* having more properties going for them, while *ajatella* and *harkita* have more properties against their occurrence, relatively speaking. Particularly striking are the odds for GENERIC and AGREEMENT types of MANNER, which increase the chances of finding *ajatella* substantially, at ratios of 23:1 and 16:1. This is due to the almost exclusive use of *ajatella* in conjunction with these specific subtypes of MANNER.

3.5 Lessons drawn: Less is more and more is less

The accuracy rate seems to reach a ceiling at around 64.6% (rising only to 65.8% with an extended variable set) for the 4-way choice for the Finnish THINK verbs and at about 52% for a 6-way choice for the Russian TRY verbs; these are in effect quite similar results, given that we are comparing a 4-way with a 6-way choice where more outcome options in general tax the level of accuracy.

Furthermore, the prediction rate appears indifferent to whether some individual variables are left out, which points in the direction of quite robust results; it is probably also a sign of high intercorrelation between contextual predictors, since it is apparent that only a subset of all theoretically possible combinations of the properties are actually ever used in practice.[16]

16 In any case, when the number of nominal variables is quite large, as is the case here, it is probable that some intercorrelation remains among the properties which can never be fully purged, irrespective of whether that would in fact be desirable (as is discussed in Footnote 15 in Section 3.4). Nevertheless, it has also been observed that variables which correlate do not necessarily substantially diminish the explanatory power of the model, as long as the correlation is not limited only to the observed data but is sufficiently general to exist also in unseen, new data (Harrell 2001: 65).

Table 8: Verb specific odds per property for all six Russian verbs

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	**3.4:1**	1:1.6	(1:1.1)	(1:1)	(1:1.2)	(1:1)
FINITE.ASPECT_PERFECTIVE	**29:1**	(1.1:1)	(1.1)	(1:4.9e7)	(1:1.1e8)	(1:3.0e7)
FINITE.MOOD_GERUND	1:8.3	(1.2:1)	**2.2:1**	**7:1**	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)	**2.4:1**	1:2	**2.1:1**	1:3.3	3.3:1
INFINITIVE.ASPECT_ IMPERFECTIVE	**6.1:1**	1:2.7	**4:1**	1:10	1:2.9	(1:1)
INFINITIVE.CONTROL_HIGH	(1:1.2)	**3.1:1**	**1.6:1**	1:6.4	**2.6:1**	**4.7:1**
INFINITIVE.SEM_ COMMUNICATION	**2.1:1**	1:1.9	(1:1.6)	(1:1)	(1.2:1)	**8.4:1**
INFINITIVE.SEM_EXCHANGE	(1.4:1)	(1:1.9)	(1:1.5)	1:11	**7.7:1**	**9.1:1**
INFINITIVE.SEM_METAPH . . . _MOTION	(1.5:1)	(1:1)	(1:1.5)	1:3.7	**6.1:1**	(1.9:1)
INF . . . . SEM_METAPH . . . _PHYS . . . _EXCH . . .	(1:1.3)	1:2.6	(1.8:1)	1:3	**4:1**	(4:1)
INF . . . . SEM_METAPH . . . _PHYS . . . _OTHER	(1.3:1)	(1:1.3)	(1:1.1)	(1:1.3)	**2.7:1**	(1.3:1)
INFINITIVE.SEM_MOTION	(1.7:1)	1:4.2	1:3.2	1:4.5	**8.1:1**	**19:1**
INFINITIVE.SEM_MOTION_ OTHER	(2.6:1)	(1:1.5)	1:3.6	(1:1.3)	**4.5:1**	**5.1:1**
INFINITIVE.SEM_PHYSICAL	**3.9:1**	**1:4.1**	(1:1.8)	(1.1:1)	**6:1**	(1.6:1)
INFINITIVE.SEM_ PHYSICAL_OTHER	**2.5:1**	(1:1.5)	1:2.1	1:2.6	**6.1:1**	**3.1:1**
SENTENCE.DECLARATIVE	1:2.8	(1:1.1)	**2.8:1**	(3.2:1)	(1:1)	(1.3:1)
SUBJECT.SEM_ANIMATE_ HUMAN	(1.5:1)	(1.4:1)	**2.5:1**	(1:1:1)	1:4	**4.1:1**

Table 9: Verb specific odds per property for all four Finnish verbs

Feature/Lexeme	ajatella	miettiä	pohtia	harkita
AGENT + GROUP	**1:5**	**1:1.9**	**4.2:1**	(1.1:1)
AGENT + INDIVIDUAL	(1:1.2)	(1:1)	(1.6:1)	(1:1.5)
CONDITION	**1:2.2**	(1.2:1)	(1:1.7)	**2.9:1**
CO-ORDINATED VERB	**1:2.1**	**2.3:1**	(1:1.2)	(1:1.2)
DURATION	**1:8.4**	**3.4:1**	(1.3:1)	(1:1)
FREQUENCY	**1:2.6**	**1.7:1**	(1:1.3)	(1.7:1)
GOAL	**3.8:1**	(1:1.8)	(1:1.8)	**1:4.7**
LOCATION	**1:3.9**	(1:1.1)	**3.7:1**	**1:2.2**
PATIENT + että ('that' clause)	**2.6:1**	**1:1.9**	**1:2**	**1:4**
MANNER + AGREEMENT	**16:1**	**1:14**	**1:4.5**	(1:7e6)
MANNER + FRAME	**2.4:1**	**1:3.6**	(1.3:1)	**1:3.8**
MANNER + GENERIC	**23:1**	**1:6.8**	(1:5e6)	(1:9e6)
MANNER + JOINT	**1:2.7**	**2.1:1**	(1:1.3)	(1.5:1)
manner + NEGATIVE	**4:1**	(1:1.8)	**1:4.6**	(1:1.7)
MANNER + POSITIVE	(1:1.4)	(1:1)	(1:1.2)	**1.8:1**
META(-COMMENT)	(1:1.2)	(1:1)	(1:1.2)	**1.6:1**
PATIENT + DIRECT_QUOTE	**1:75**	**3:1**	**8.1:1**	(1:8.1e6)
PATIENT + INDIRECT_QUESTION	**1:14**	**4.2:1**	**2.8:1**	(1:1.2)
PATIENT + INFINITIVE	**5.3:1**	(1:4e6)	(1:4.7)	(1.4:1)
PATIENT + PARTICIPLE	**5.3:1**	(1:4e6)	(1:3.3)	(1.1:1)
PATIENT + ABSTRACTION	**1:4.1**	**1.5:1**	**4.1:1**	(1:1)
PATIENT + ACTIVITY	**1:7.1**	(1:1.3)	**1.6:1**	**9:1**
PATIENT + COMMUNICATION	**1:9.6**	**2.8:1**	**3:1**	(1.8:1)
PATIENT + EVENT	(1.4:1)	(1:1)	(1:1)	(1:3)
PATIENT. INDIV . . ./GROUP	**2.7:1**	**1:1.9**	**1:3.4**	(1:1.2)

Table 9 (Cont.)

Feature/Lexeme	ajatella	miettiä	pohtia	harkita
QUANTITY	(1:1.5)	**2.6:1**	(1:1.3)	**1:3**
REASON/PURPOSE	**1:2.3**	(1.1:1)	(1.3)	(1.6:1)
SOURCE	**3.1:1**	(1:1.3)	**1:3.5**	**1:7.5**
TIME + DEFINITE	**1:2.5**	(1:1)	**2.3:1**	(1:1.3)
TIME + INDEFINITE	**1:1.7**	**1.5:1**	(1:1)	(1.2:1)
VERB-CHAIN + ACCIDENTAL	**5.6:1**	(1:2.3)	(1:2.1)	(1:1e7)
VERB-CHAIN + EXTERNAL	**2.5:1**	(1:1.3)	(1:1.4)	(1:1.1)
VERB-CHAIN + NECESSITY	**1:2.9**	**2:1**	(1:1)	(1.4:1)
VERB-CHAIN + POSSIBILITY	(1.2:1)	(1.1:1)	(1:1.2)	(1.2:1)
VERB-CHAIN + TEMPORAL	**1:3.8**	**1.8:1**	**2.4:1**	**1:6.5**
VERB-CHAIN + VOLITION	(1:1.6)	(1.6:1)	(1:1)	(1:1.6)
COVERT (AGENT/SUBJECT)	(1.1:1)	(1.2:1)	(1:1.3)	(1:1.3)
FIRST (PERSON)	(1:1.2)	(1.8:1)	**1:3.5**	(1.9:1)
SECOND (PERSON)	(1:1.5)	**2.4:1**	**1:2.4**	(1:1.5)
THIRD (PERSON)	(1:1.6)	(1.3:1)	(1:1)	(1.6:1)
INDICATIVE (MOOD)	**2:1**	(1:1.5)	(1:1.2)	(1:1.2)
CONDITIONAL (MOOD)	(1.3:1)	**1:1.9**	(1:1.4)	**2.3:1**
NEGATION (POLARITY)	**2.1:1**	(1:1.4)	**1:2.1**	(1.1:1)
PASSIVE (VOICE)	(1:1.6)	(1:1.1)	**1.9:1**	(1.1:1)
PLURAL (NUMBER)	(1.1:1)	**1:1.7**	**1.6:1**	(1.2:1)
CLAUSE-EQUIVALENT	(1.1:1)	(1:1.7)	(1:1.1)	(2:1)

At the same time, in 48% of all Russian cases and 33.4% of all Finnish cases (i.e. 100% – accuracy) a verb other than the one originally used is predicted (which does not necessarily mean it would be wrong, see Arppe and Divjak In preparation). Moreover, in 52.6% of the Russian cases and 24.2% of the Finnish cases no particular verb is strongly preferred (i.e. max[P(*Verb*|*Context*)] < 0.5, in which case no verb by itself is allotted the majority of the available probability). These differences are not surprising given that the Russian near-synonyms form a homogeneous group with regard to argument structure, while the Finnish near-synonym set allows variation in the argument structure. Could this result also signal that we have we missed something important in the linguistic analysis schemes used to annotate the datasets? We find it difficult to identify any additional contextual properties or new property clusters pertaining to current, conventional models of morphology, syntax and semantics, and applicable within the immediate sentential context that are not incorporated in the current analysis in one way or another. For Finnish, were we to have even more data, we might want to include in the multivariate analysis rarer semantic subclassifications of the syntactic argument types, which were excluded due to the constraints on the number of variables in the model. In the case of Russian, we would need more data to model all available properties reliably, as well as include infrequent optional elements such as adverbs and other non-obligatory arguments.

More in particular, choices in equiprobable cases (Arppe and Divjak In preparation) cannot always be explained on the basis of the combined preferred properties alone (something we will elaborate on in Section 4.3) but require us to look at other properties as well (keeping in mind that some properties were excluded from the dataset for modeling purposes, see Section 3.1). Yet even if near-synonyms are equiprobable in a given context, each lexeme still presents a different perspective on the situation. The semantic differences between any of the TRY or the THINK lexemes seem embedded and manifested in the lexemes themselves and would not necessarily have or require explicit manifestation. On a cognitive linguistic, usage-based view, this might be due to the fact that every encounter with a verb contributes to a "cloud" of exemplars from which salient properties are extracted. These properties become part of the lexeme-specific information for each verb – be it called Idealized Cognitive Model or prototypical usage – and are inferred even when the salient contextual properties that initially triggered them are not available. However, implementations of this theoretically possible process are absent from the cognitive linguistic literature.

In the next section we demonstrate how this continuum from individual exemplars to abstracted prototypes can be observed in corpus data and how, using statistical modeling techniques, salient properties can be extracted from a

tagged cloud of exemplars. To this end, we build upon the results presented in this section.

4 Extremes of a continuum: the exemplar and the prototype model

As mentioned at the outset of this paper, two distinct views about how linguistic categories are stored and represented as cognitive structures dominate the categorisation-by-similarity scene, namely the *prototype* and *exemplar* theories. In accordance with the single prototype view, a category/concept is assumed to be stored as a highly abstract representation, consisting of an aggregate of properties that are characteristic of the concept (e.g. Rosch 1973) – rather than strictly defining and delineating it. In contrast, according to the full exemplar view, a category/concept is presumed to be represented as detailed memory traces of all the individually encountered exemplars of the concept, with little or no abstraction taking place across these stored exemplars (e.g. Hintzman 1986; Nosofsky 1986).

With Vanpaemel and Storms (2008, 2010) we would rather see the exemplar and prototype routes as opposite ends along a continuum. This perspective is in line with a usage-based view on language acquisition: on a usage-based view, prototypes would emerge from repeated exposure to and abstraction over exemplars.[17] The multivariate analysis technique, introduced in Section 3, has two key attractive characteristics as stepping stones towards showing how a mixed representation model can be achieved on the basis of usage data as recorded in corpora. We will illustrate this procedure using data from Russian.

4.1 Property-wise verb-specific odds

As discussed in Section 3.4, a model created with polytomous logistic regression (using the one-vs-rest heuristic) provides probability estimates for the (proportional) occurrence of an outcome, in this case a verb within a synonym set, given the contextual occurrence of a combination of linguistic properties incorporated in the model, using the odds assigned to the properties in question as a

17 The categorization literature reports both exemplar-to-prototype shifts and prototype-to-exemplar shifts. This difference seems to be caused by the concepts used: those studies finding exemplar-to-prototype shifts used concepts with defining features while the studies finding prototype-to-exemplar shifts did not (Beatu and Schulz 2010).

result of fitting the model to the data. For example, the probability of encountering *probovat'* given a context consisting of the set of properties listed under (12) is 0.92.

(12) P(*probovat'*|{CLAUSE.MAIN, FINITE.ASPECT_PERFECTIVE, INFINITIVE.CONTROL_HIGH, INFINITIVE.SEM_MOTION, SUBJECT.SEM_ANIMATE_HUMAN}) = 0.92

Попоробуй/poprobuj уйти отсюда или отказаться от обещанного! [Č. Ajtmatov. Belyj paraxod. #252]

"Try [probovat'] to go away from here or to renounce what was promised."

Secondly, the *one-vs-rest* heuristic can be understood to highlight those properties which distinguish the individual outcome classes (in this case the near-synonymous verbs) from all the rest (within the same set), in natural terms as *odds* (see Table 8 above). As an example case, take the Russian *probovat'*, for which the property-wise odds are repeated here in Table 10. Significant odds are indicated in bold-face while non-significant odds are parenthesized; fractions with denominator 1 (in the table displayed as X:1) signal odds in favour while fractions with numerator 1 (in the table displayed as 1:X) signal odds against. Individual odds (parameter values) which are greater than 1.0 for some property and *probovat'* can be interpreted to reflect the increased chances of occurrence of *probovat'* when the property in question is present in the context. Conversely, odds less than 1.0 denote a decreased chance of the occurrence of *probovat'* in such a context. These odds are abstract representations of the association strength between a verb and a property in the mental lexicon. For the average speaker or hearer, they are an estimation of what they can expect to hear or say next.

4.2 The exemplar route: ranking individual sentences

Important for the current study is the fact that the probability estimates can also be used to rank individual sentences (and the combinations of contextual properties they incorporate) included in the training corpus in terms of how "exemplary" they are with respect to the TRY verb they contain. Gries (2003b) is an early example of how probability estimates based on scores from a multifactorial analysis of corpus data embodying speaker choices can be used to sort instances of constructions according to their degree of prototypicality. However, simply picking out the sentences with the highest probability estimates for each verb would yield

Table 10: Odds for contextual properties in conjunction with *probovat'*

Property/Verb	*Probovat'*
(Intercept)	1:22
CLAUSE.MAIN	**3.4:1**
FINITE.ASPECT_PERFECTIVE	**29:1**
FINITE.MOOD_GERUND	**1:8.3**
FINITE.MOOD_INDICATIVE	**1:2.8**
FINITE.TENSE_PAST	(1:1)
INFINITIVE.ASPECT_IMPERFECTIVE	**6.1:1**
INFINITIVE.CONTROL_HIGH	(1:1.2)
INFINITIVE.SEM_COMMUNICATION	**2.1:1**
INFINITIVE.SEM_EXCHANGE	(1.4:1)
INFINITIVE.SEM_METAPHORICAL_MOTION	(1.5:1)
INFINITIVE.SEM_METAPHORICAL_PHYSICAL_EXCHANGE	(1:1.3)
INFINITIVE.SEM_METAPHORICAL_PHYSICAL_OTHER	(1.3:1)
INFINITIVE.SEM_MOTION	(1.7:1)
INFINITIVE.SEM_MOTION_OTHER	(2.6:1)
INFINITIVE.SEM_PHYSICAL	**3.9:1**
INFINITIVE.SEM_PHYSICAL_OTHER	**2.5:1**
SENTENCE.DECLARATIVE	**1:2.8**
SUBJECT.SEM_ANIMATE_HUMAN	(1.5:1)

too many exemplars because similar contexts will receive similar probability estimates. For instance, the 248 occurrences of *probovat'* correspond to only 100 distinct property combinations (in terms of the properties included in the multivariate model). Thus, such straightforward selection, based merely on sentence-wise probability estimates, will lead to examples which are essentially duplicates of the same contexts and properties.

We may assume that for each verb there are several typical usage contexts, which cannot necessarily be reduced to one observed sentence that would

aggregate them all. For example, the infinitive that is attempted can only have one semantic characterization in any particular sentence. Nevertheless, there probably will be substantial overlap in context as the verbs individually and as a group do share some common contextual properties, for example, they typically – and overwhelmingly – enlist HUMAN BEINGS as AGENTS.

Since the entire set of sentences in the data set is classified according to 18 contextual properties selected in the model (see Section 3.1), we can use the available sentence-wise property sets as input for a statistical clustering algorithm to sort the underlying sentences into groups which are internally similar but group-wise distinct in terms of their constituent properties. Hierarchical agglomerative clustering (HAC), which has been successfully applied in the study of the selected TRY verbs (Divjak 2004; Divjak and Gries 2006), is an attractive clustering technique that allows us to group the annotated exemplars into groups or clusters of highly similar items (for a general introduction to cluster analysis for linguists, see Johnson (2008: Chapter 6), Baayen (2008: Chapter 5), Gries (2009: Chapter 5, Section 5), Divjak and Fieller (Forthcoming). Since the sentence-wise properties (as logical true/false variables) are binary, the corresponding *Binary* distance measure seems most appropriate, as is also the case for the *Ward* clustering algorithm, which aims at finding compact, spherical clusters.

The type of clustering technique applied here is based on ideas presented in Arppe (2008: 248–252) and lets us extract whatever number of clusters we deem appropriate for our purpose, i.e. for discovering property clusters. Setting the number of distinct property clusters is a result of trial and error.[18] We have opted for the smallest number of clusters that ensures each of the six Russian TRY verbs is allotted at least one exemplary sentence (with the highest probability for the property cluster in question). Each individual property cluster typically represents many individual example sentences in the data, but the full property combinations of such sentences are not necessarily entirely identical, nor are the verb-specific probability estimates. With the overall number of property clusters

18 Although there are many methods available for selecting the "ideal" number clusters, we opted for "trial and error" because the combination of 1) clustering the entire data based on the contextual predictors excluding the outcomes and 2) probability estimates for the outcomes provided by polytomous logistic regression, which are together used to extract a subset of exemplary sentences, favors the outcome with the greatest variety in terms of the contexts it occurs in over the more focused outcome. From each contextually determined subset of the data (i.e. cluster) only one sentence is selected, which is among those with the highest estimated probability for the verb that actually occurs in the original sentences (within the cluster in question). Manual trial of setting the exemplar set size is necessary to ensure that each outcome gets at least one such exemplar sentence, and to explore whether a somewhat larger exemplar set increases the number of exemplar sentences even for the rarer outcomes.

set to 40,[19] we can extract – instead of the original 248 example sentences for *probovat'* – a substantially smaller subset of 13 exemplars for this verb (see Table 11 for the contextual property combinations used in these examples). The number of such distinct property clusters for the others TRY verbs varies substantially, being 12 for *silit'sja*, 10 for *norovit'*, 3 for *starat'sja*, and only one each for both *pytat'sja* and *poryvat'sja*. This variation among the verbs can be seen as a manifestation of the diversity of contexts in which each verb typically occurs. Crucially, the partitioning of exemplars into categories is achieved by pitting a certain category against the others. This ensures that properties are given a boost if they co-occur more often with the category of interest and less often with the remaining categories.

Cluster 3 in Table 11, for example, tells us that given the context of {CLAUSE.MAIN, FINITE.ASPECT_PERFECTIVE, FINITE.MOOD_INFINITIVE, INFINITIVE.CONTROL_HIGH, INFINITIVE.SEM_PHYSICAL_OTHER} we are most likely to find *probovat'*. P(*probovat'*|Context) is the highest for all TRY verbs in that particular context, indeed. An example of a sentence incorporating this property combination with a probability estimate of 0.771 for *probovat'* is given in (13):

(13) Вы меня на крыше подстрахуете, я спущусь в окно, попробую/ **poprobuju** открыть сейф – как? Это легавых не касается. Давай фотоаппарат. [F. Neznanskij. Operacija "Faust". #374] [P(*probovat'*|Context) = 0.771

"You cover me from the roof, I'll go down through the window, I'll try [probovat'] to open the safe. How? That's none a of police spy's business. Give me the camera."

Since these sentences are exemplary only in comparison with the other sentences, we provide an exemplary exemplar for each of the six Russian TRY verbs below,

19 For the Finnish THINK verbs, setting the number of clusters at 100, reflecting the larger size of the dataset and greater number of possible contextual properties in comparison to the Russian one, we can extract from the entire dataset of 3404 sentences a subset of 100 exemplars, of which predominantly 71 are with *ajatella*, 9 with *miettiä*, 16 with *pohtia*, and 4 with *harkita*, again reflecting the diversity of contexts associated with each verb observed generally in Arppe (2008). Note that it could well be desirable to extract these exemplars in roughly more equal proportions than shown in this paper or to emphasize other characteristics in the exemplar sets, and Arppe (2008: 248–252) suggests some strategies with which the selection process could be tweaked to achieve this, e.g. by giving more weight to contextually richer exemplars that incorporate larger numbers of contextual properties or by using probability bins.

Table 11: The 13 distinct property clusters for *probovat'* ('+' indicates inclusion of property in an individual cluster)

Property/Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13
CLAUSE.MAIN	+	+	+	+	–	+	+	+	+	+	+	+	+
FINITE.ASPECT_PERFECTIVE	+	+	+	–	+	+	+	+	+	+	+	+	+
FINITE.MOOD_INDICATIVE	+	–	+	+	–	+	+	–	–	–	–	+	–
FINITE.TENSE_PAST	+	–	–	+	–	+	+	–	–	–	–	+	–
INFINITIVE.ASPECT_IMPERFECTIVE	–	–	–	+	–	+	+	–	–	–	–	+	–
INFINITIVE.CONTROL_HIGH	+	+	+	+	+	+	+	–	–	+	–	–	–
INFINITIVE.SEM_COMMUNICATION	+	–	–	–	–	+	–	–	–	–	–	–	–
INFINITIVE.SEM_EXCHANGE	–	–	–	–	–	–	–	–	+	–	–	–	–
INFINITIVE.SEM_MOTION	–	–	–	–	+	–	–	–	–	+	–	+	–
INFINITIVE.SEM_MOTION_OTHER	–	+	–	–	–	–	–	–	–	–	–	–	–
INFINITIVE.SEM_PHYSICAL	–	–	–	+	–	–	–	+	–	–	–	–	–
INFINITIVE.SEM_PHYSICAL_OTHER	–	–	+	–	–	–	+	–	–	–	+	–	–
SENTENCE.DECLARATIVE	+	+	–	–	+	–	+	+	+	–	–	+	–
SUBJECT.SEM_ANIMATE_HUMAN	+	+	+	+	+	+	+	+	+	+	+	+	+

i.e. sentences that combine a distinct property combination with a high probability estimate:

(14) Но такие ребяческие желания могли показаться серьезным людям смешными. Опасаясь этого, он всеми силами пытался/**pytalsja** не выдавать себя. Но это не совсем удавалось. Трудно было ему скрыть свое счастье – горячий румянец отчетливо проступал на смуглых крепких щеках. [Č. Ajtmatov. Pegij pes, beguščij kraem morja Vladimiru Sangi. #857 [P(*pytat'sja*) = 0.41]

"But such childish desires might seem ridiculous to serious people. Therefore, he tried [pytat'sja] as hard as he could not to give himself away. But he didn't quite manage. It was difficult for him to keep his happiness hidden – a burning blush appeared distinctly on his firm swarthy cheeks."

(15) Только интересно, сижу и соображаю, что я раньше почувствую, «пиф-паф» или удар в затылок? Соображаю и стараюсь/**starajus'** убить в себе нерв жизни, чтобы ничего не вспоминать, не сопливиться, чтобы ни о чем не жалеть, никого не хаять и никого не любить. [Ju. Aleškovskij. Kenguru. #1119, P(*starat'sja*) = 0.33]

"[. . .] I am thinking about this and am trying [starat'sja] to kill the pulse of life in myself, so as to reminisce no more, to whine no more, not to regret anything, not to pick on anyone and not to love anyone."

(16) Насчет бабочки, Эмма, какая-то ерунда, – заговорил Никитин пасмурно, тщетно силясь/**siljas'** найти немецкие слова.
– Не в этом дело. А, черт, язык! Ну, как же тебе объяснить? [Ju. V. Bondarev. Bereg. #577, P(*silit'sja*) = 0.66]

"As far as the butterfly is concerned, Emma, that's nonsense, said Nikitin gloomily, vainly trying [silit'sja] to find the German words. That's not what this is about. Ah, damn, this language! Well, how shall I explain it to you?"

(17) – Хорошо. Он учится во вторую смену, с двенадцати до семи. К восьми вечера он будет здесь. Настя отправилась к себе, с трудом удерживая норовящие/**norovjaščie** выскользнуть из рук папки и с удивлением думая о том, почему это она так спокойно позволяет Заточному распоряжаться ее временем. [A. B. Marinina. Nastja Kamenskaja. #40. P(*norovit'*) = 0.81]

"Ok. He is in the second group, and goes to class from 12 to 7. He'll be here about 8. Nastja went back, barely managing to carry the folders that were trying [norovit'] to slide out of her hands, and wondering why she so calmly let Zatočnyj take up her time."

(18) Потом мне рассказывали что и раньше он несколько раз порывался/**poryvalsja** уйти из семьи. Но во-первых уйти было некуда а во-вторых жена приходила жаловаться. [F. Iskander. Sozvedenie Kozlotura. #743. P(*poryvat'sja*) = 0.39]

"Later they told me that he had already tried [poryvat'sja] to leave his family on a couple of occasions before. But, first of all, he had nowhere to go, and second, his wife came and complained."

For the Finnish THINK verbs more than twice as much data was available in a simpler synonymy-setting and a set of 100 overall similar exemplary exemplars

was selected (cf. Arppe 2008: 248–252). What we now have are substantially reduced sets of sentences (40 out of 1351 for the Russian TRY verbs and 100 out of 3404 for the Finnish THINK verbs) which represent both the data sets as a whole as well as the TRY (or THINK) verbs which are most typically used in such exemplary contexts. This process resembles Verbeemen et al.'s (2007) gradual abstraction process. Verbeemen et al. (2007) make a partition of the category exemplars for each category involved and then construct, for every subset of exemplars, a prototype by averaging over all the exemplars on that subset; we employ estimated probability to yield the most prototypical exemplar per subset (i.e. cluster, the set of which is determined by distribution of properties over all data).

Next, we start from the other end and look at which sets of properties are overall typical and characteristic for each individual verb, and what types of general abstract characterizations, i.e. prototypes, we can deduce for each verb on the basis of such characteristic property sets.

4.3 The prototype route: aggregating properties as a prototype

The verb-specific odds for the analytical linguistic properties selected in the model – which are per definition generalizations abstracted over the individual words and phrases in the data – can be aggregated to construct an abstraction which, as a whole, embodies and represents the prototype of each verb, when contrasted with the rest of the verbs in the near-synonym set.

Thus for the Russian TRY verbs, out of a total of 1,351 individual example sentences and the property combinations they incorporate, 660 distinct combinations of a verb plus a context type can be distinguished. If we ignore the outcome verb, this number reduces further to 296 unique property combination types. And if we select only permissible property combinations with significantly favorable odds, we are left with only 20. This is an excellent example of different intermediate levels of (prototype) extraction. For e.g. *probovat'*, the three such combinations of allowable properties with significant odds in favor of the verb are given in (19):

(19)

a. {CLAUSE.MAIN, FINITE.ASPECT_PERFECTIVE, INFINITIVE.ASPECT_IMPERFECTIVE, INFINITIVE.SEM_COMMUNICATION}
b. {CLAUSE.MAIN, FINITE.ASPECT_PERFECTIVE, INFINITIVE.ASPECT_IMPERFECTIVE, INFINITIVE.SEM_PHYSICAL}
c. {CLAUSE.MAIN, FINITE.ASPECT_PERFECTIVE, INFINITIVE.ASPECT_IMPERFECTIVE, INFINITIVE.SEM_PHYSICAL_OTHER}

Ultimately, these altogether 20 permissible property combinations (for all six TRY verbs) with significantly favourable odds reduce to as few aggregates of properties with strongly favourable odds as there are verbs, i.e. six. Let us illustrate this on the basis of the Russian verb *probovat'*.

We can construct the aggregate of contextual properties associated with *probovat'* using those properties which have been deemed – following statistical analysis of the entire data set – to have significant odds in favour of *probovat'*, boldfaced in Table 11 and repeated here under (20). Note that only one of the three semantic characterizations for the infinitive can be instantiated within one and the same verb. Thus, the aggregate of properties represents in fact three permissible property combinations, listed above under (19) as a, b and c.

(20)

Odds	Property
3.4:1	CLAUSE.MAIN
29:1	FINITE.ASPECT_PERFECTIVE
5.4:1	INFINITIVE.ASPECT_IMPERFECTIVE
2.1:1	INFINITIVE.SEM_COMMUNICATION
3.9:1	INFINITIVE.SEM_PHYSICAL
2.5:1	INFINITIVE.SEM_PHYSICAL_OTHER

At this point the question arises: are these properties, as a whole, manifestations of the core of a prototype for *probovat'*, i.e. the set of properties that tips the scale significantly in favour of a verb but does not exhaustively define it (if only because not all conceivable properties are represented in our tagset)? When comparing these prototypes to the ones previously presented in Divjak (2004; 2010) or Divjak and Gries (2006) that were obtained by applying exploratory cluster analysis to a variable set that does include optional elements such as adverbs, it emerges that the ones featured here represent the cores of the prototypes indeed.

It is plausible to interpret – in a way reminiscent of the theorizer's mental leap taken while trying to make overall sense of a number of interpersonally observable yet possibly heterogeneous facts concerning a phenomenon (Dennett 1991) – the properties listed under (20) as conveying the notion of telling someone to *try* (using the perfective aspect hence signalling the *attempt* should be taken to its natural conclusion and with limitations imposed on the time or effort invested), and carry out a physical action, to manipulate someone or something, or to communicate (using the imperfective, i.e. without insisting that the attempted *action* be taken to its natural end). This interpretation of *probovat'* explains why this verb is typically characterized as an "experimental attempt" (Apresjan et al. 1999; Divjak and Gries 2006: 18; Divjak 2010: 164, 169), and why it

is the most frequently used TRY verb in mother-child interaction (Stoll corpus, see Divjak and Gries 2006).

The preceding sections have demonstrated that properties such as those listed under (20), which we argue make up the core of the prototype, have been distilled from the individual example sentences contained in the analyzed dataset; this closes the circle. What further strengthens our argument for considering full exemplar and single prototype models of classification as the opposite ends of one and the same continuum of abstraction is the fact that the most exemplary sentences incorporate the most typical properties. Exemplary exemplars and property combinations with significant odds map well onto each other, although not perfectly, as the exemplars also incorporate the effects of properties not included in the abstract prototypes.[20] If we look at which verb is used in the property combinations listed under (21), we see that in the sole sentence with property combination (21a) *probovat'* was used, as was the case in 3 out of 3 sentences with property combination (19b) and 3 out of 3 with property combination (19c) (see 21). That is, all of the sentences with the prototypical property combinations have as the actually occurring verb the one associated with the prototype.

(21)

a. **1/1:** {CLAUSE.MAIN, FINITE.ASPECT_PERFECTIVE, INFINITIVE.ASPECT_IMPERFECTIVE, INFINITIVE.SEM_COMMUNICATION}
b. **3/3**: {CLAUSE.MAIN, FINITE.ASPECT_PERFECTIVE, INFINITIVE.ASPECT_IMPERFECTIVE, INFINITIVE.SEM_PHYSICAL}
c. **3/3**: {CLAUSE.MAIN, FINITE.ASPECT_PERFECTIVE, INFINITIVE.ASPECT_IMPERFECTIVE, INFINITIVE.SEM_PHYSICAL_OTHER}

For the other Russian TRY verbs, the prototype cores suggested by the analysis are given in (22) through (26). *Pytat'sja* in (22) has very few core properties; an ideal *pytat'sja* context merely requires the attempt to be located in the past and to

20 It holds at the most abstract level that exemplars have high probabilities because they incorporate contextual properties which have one or more highly significant odds in favour of the verb in question. Thus, properties with high odds for a particular verb can be expected to occur in sentences that receive a high probability estimate. Yet we do not use the verb-specific aggregated properties as a starting point for selecting exemplars. Rather, we look at the properties in the entire data set, and pick sentences (and the verbs they contain) that best exemplify primarily the entire data set, and only secondarily the individual verbs. Both the exemplary exemplars and the aggregated property prototypes result from applying one and the same statistical method to exactly the same data, but they follow distinct paths: the exemplars take into consideration also other properties than those incorporated in the aggregated prototype, while only properties with significant odds in favour of a particular verb are considered for the aggregated prototype.

be aimed at carrying out an action over which the subject has control. This is hardly surprising given that, as argued in Divjak (2010: 162–165), *pytat'sja* is the most neutral type of attempt fitting virtually every situation.

(22) Pytat'sja

Odds	Property
2.4:1	FINITE.TENSE_PAST
3.1:1	INFINITIVE.CONTROL_HIGH

Starat'sja in (23) is more precise in requiring a human being to undertake an attempt directed at carrying out an action that is ongoing or happens repeatedly (imperfective) over which s/he has high control. The TRY verb itself takes the form of a gerund thus relating the attempt to carry out a controllable action to the situation expressed in the main clause with respect to time, cause, condition, reason etc. (cf. Divjak 2010: 212). The entire situation is typically reported on by means of a declarative clause.

(23) Starat'sja

Odds	Property
2.2:1	FINITE.MOOD_GERUND
4:1	INFINITIVE.ASPECT_IMPERFECTIVE
1.6:1	INFINITIVE.CONTROL_HIGH
2.8:1	SENTENCE.DECLARATIVE
2.5:1	SUBJECT.SEM_ANIMATE_HUMAN

Silit'sja in (24) requires contexts in which the attempt is located in the past (since it is difficult if not impossible to ascertain the infelicity of an attempt while it is ongoing), presented as a gerund, and aimed at accomplishing something over which the subject has little or no control.

(24) Silit'sja

Odds	Property
7:1	FINITE.MOOD_GERUND
2.1:1	FINITE.TENSE_PAST
10:1	INFINITIVE.ASPECT_PERFECTIVE (complementary case of INFINITIVE.ASPECT_IMPERFECTIVE)
6.4:1	(INFINITIVE_CONTROL_MEDIUM/LOW: complementary case of INFINITIVE_CONTROL_HIGH)

The last two verbs, *norovit'* and *poryvat'sja*, pose much more precise (but also more diverse) requirements on the context they find ideal, in particular on the type of action they like to be aimed towards. *Norovit'* in (25) prefers situations in which non-human subjects carry out (perfective) an action that is either physical action or motion, possibly aimed at a person or object or happening between two people. Some of the base-actions, i.e. motion and exchange, can be encountered in their metaphorical use as well.

(25) Norovit'

Odds	Property
6:1	INFINITIVE.ASPECT_PERFECTIVE (complementary case of INFINITIVE.ASPECT_IMPERFECTIVE)
2.6:1	INFINITIVE.CONTROL_HIGH
7.7:1	INFINITIVE.SEM_EXCHANGE
6.1:1	INFINITIVE.SEM_METAPHORICAL_MOTION
4:1	INFINITIVE.SEM_METAPHORICAL_PHYSICAL_EXCHANGE
2.7:1	INFINITIVE.SEM_METAPHORICAL_PHYSICAL_OTHER
8.1:1	INFINITIVE.SEM_MOTION
4.5:1	INFINITIVE.SEM_MOTION_OTHER
6:1	INFINITIVE.SEM_PHYSICAL
6.1:1	INFINITIVE.SEM_PHYSICAL_OTHER
4:1	SUBJECT.SEM_NON_HUMAN (complementary case of SUBJECT.SEM_ANIMATE_HUMAN)

Poryvat'sja in (26), finally, prefers contexts in which human subjects have undertaken an attempt to carry out an action over which they have control, such as communicating a message, exchanging something, undertaking a physical effort or moving themselves or others.

(26) Poryvat'sja

Odds	Property
3.3:1	FINITE.TENSE_PAST
4.7:1	INFINITIVE.CONTROL_HIGH
8.4:1	INFINITIVE.SEM_COMMUNICATION
9.1:1	INFINITIVE.SEM_EXCHANGE
19:1	INFINITIVE.SEM_MOTION
5.1:1	INFINITIVE.SEM_MOTION_OTHER
3.1:1	INFINITIVE.SEM_PHYSICAL
4.1:1	SUBJECT.SEM_ANIMATE_HUMAN

The reader who is familiar with previous work on TRY verbs in Russian will notice the similarity between the prototypes presented here and those that fell out from an experimentally validated exploratory cluster analysis (Divjak and Gries 2006; Divjak and Gries 2008). Although the difference in datasets used does not allow us to present the current results as validation of the previously obtained prototypes, the convergence signals that the prototypes of the verbs seem stable: since they emerge in different constellations and as a result of different statistical applications, they are unlikely to be a mere by-product of the statistical analysis techniques used.

The same procedure can be repeated to construe prototypes for the Finnish THINK verbs. *Ajatella* in (27) reflects a temporally indefinite, continuous *way of thinking*, whether a state of mind, intention, or opinion/stance/attitude/perspective, that may conform with or diverge from that held by others (i.e. *think* alike vs. differently) or negated. In addition, *ajatella* is individual both in its agency and its patients.

(27) Ajatella

23:1	MANNER + GENERIC
16:1	MANNER + AGREEMENT
5.6:1	VERB-CHAIN + ACCIDENTAL
5.3:1	PATIENT + INFINITIVE
5.3:1	PATIENT + PARTICIPLE
4:1	MANNER + NEGATIVE
3.8:1	GOAL
3.1:1	SOURCE
2.7:1	PATIENT + INDIVIDUAL/GROUP
2.6:1	PATIENT + että ('that' clause)
2.5:1	VERB-CHAIN + EXTERNAL
2.4:1	MANNER + FRAME
2.1:1	NEGATION (POLARITY)
2.1	INDICATIVE (MOOD)

The *thinking* process denoted by *miettiä* in (28) is temporally anchored or constrained, either with respect to its duration (brief or long), quantity (little or much), or frequency (once, twice, often), or via a temporal expression in the verb-chain indicating beginning, ending, or continuation. However, an explicit expression of time linked with *miettiä* is indefinite in nature. In terms of its agency it is clearly individual and interpersonal, specifically in its association with the second person. The patients, i.e. objects, of *miettiä* may be any form of human

communication, whether indirect questions, direct quotes, or words denoting communication.

(28) Miettiä

4.2:1	PATIENT + INDIRECT_QUESTION
3.4:1	DURATION
3:1	PATIENT + DIRECT_QUOTE
2.8:1	PATIENT + COMMUNICATION
2.6:1	QUANTITY
2.4:1	SECOND (PERSON)
2.3:1	CO-ORDINATED VERB
2.1:1	MANNER + JOINT
2:1	VERB-CHAIN + NECESSITY
1.8:1	VERB-CHAIN + TEMPORAL
1.7:1	FREQUENCY
1.5:1	PATIENT + ABSTRACTION
1.5:1	TIME + INDEFINITE

Pohtia in (29) can be interpreted as a *thinking* process undertaken by all sorts of human groups as agents – whether understood as collectives, locations referring primarily to organizations, consisting of countable individuals (plural) or impersonal, unidentified groups (represented by the Finnish passive voice). With respect to the patient or object of the *thinking* activity, *pohtia* can be construed to be prototypically linked with abstract concepts including activities and various forms of verbal communication (direct quotes, indirect questions, or expressions referring to media or acts of communication). Finally, *pohtia* as a *thinking* activity is prototypically seen to have a temporal anchor, whether expressed by a temporal auxiliary verb indicating beginning, continuation, or ending the *thinking* process, or an argument referring a definite moment in time when the *thinking* activity has taken place.

(29) Pohtia

Odds	Property
4.2:1	AGENT + GROUP
3.7:1	LOCATION
1.6.1	PLURAL
1.9:1	PASSIVE
8.1:1	PATIENT + DIRECT_QUOTE
4.1:1	PATIENT + ABSTRACTION
3:1:1	PATIENT + COMMUNICATION

2.8:1 PATIENT + INDIRECT_QUESTION
1.6:1 PATIENT + ACTIVITY
2.4:1 VERB-CHAIN + TEMPORAL
2.3:1 TIME-POSITION + DEFINITE

Finally, *harkita* in (30) is specifically associated with an act/action as its patient or object, which orients this *thinking* process towards future. This resulting action may be hedged in various ways, with an explicit expression of a condition, conditional mood, or a meta-comment (prototypically *ehkä* 'maybe'). Moreover, *harkita* is also associated with thoroughness. In fact, *harkita* is the only one of the four Finnish THINK verbs for which it would be, at least in principle, possible to observe all significant favourable properties within one single sentence, incorporated in the constructed sentence *Ehkä* ***harkitsisin*** *tarkkaan lopettamista, jos* . . . 'I maybe would carefully **consider** quitting, if . . .'.

(30) Harkita
9:1 PATIENT + ACTIVITY
2.9:1 CONDITION
2.3:1 CONDITIONAL (MOOD)
1.8:1 MANNER + POSITIVE
1.6:1 META-COMMENT

5 Conclusion

We have demonstrated one way of systematically analyzing usage data as contained in corpora to yield a scheme, compatible with usage-based theories of language, by which the assumptions of both the prototype and exemplar theories can be operationalized. This lends support for viewing the prototype and exemplar theories as being positioned on a continuum of abstraction, instead of as being mutually exclusive or irreconcilable. Stronger even, our data underline the need to consider *varying abstraction* models: the mental lexicon could effectively embody several kinds of representations, i.e. exemplars as well as abstractions that vary in degree, and that simultaneously, since such representations can be observed in the resultant linguistic output. Exemplars are visibly present in language and a general abstraction can be arrived at on the basis of co-occurrence data available using probabilistic techniques. Given that both exemplars and a prototype are available to the speaker given the input s/he receives and the domain general cognitive learning mechanisms s/he possesses, maybe we should abandon trying to accommodate the facts one way or another and combine both

approaches? Yet since several "intermediate" abstractions are arrived at with greater ease than one overarching prototype is, we propose to not merely allow both exemplars and prototypes in a model, but to acknowledge varying degrees of abstraction in between. This would also be in line with the finding that both lower and higher level abstractions or schemas are active in representing morphological and syntactic knowledge (Verhagen 2005; Dąbrowska 2008).

Our analysis has revealed key properties of the structures under investigation, in our case near-synonymous verbs, the combinations in which these properties preferably co-occur, as well as a subset of exemplary sentences embodying these properties. This approach lets us stipulate a prototype-core emerging from the aggregate of the exemplary exemplars. Such a prototype merges information from one or more exemplary examples, instantiating (parts of) an (idealized) property configuration that occurs significantly more frequently with one structure than with one other or multiple others. We have argued that these property configurations trigger the selection of a structure when the properties are explicitly present in the context and are invoked implicitly through the selection of the structure when the properties are not evident in the context. The structures under investigation can be anything from phones or morphemes to syntactic constructions.

The fact that these results were obtained using a statistical technique that tracks frequencies and models proportions lends credibility to our representation of the linguistic phenomenon in terms of what we know about human cognitive processing. Over the past decade, numerous studies have been published supporting the claim that infants are equipped with powerful statistical language learning mechanisms (Saffran et al. 1996). If speakers do indeed model input statistically, they may well be operating with prototypes that are at least functionally similar to what the regression technique outputs. Yet it needs to be borne in mind that, in this paper, we have shown that prototypes can be extracted from exemplars by tracking probabilities in input. We are not claiming that we have modelled the way in which prototypes emerge from exemplars in reality, i.e. in learning a language in childhood or adolescence, or how they change over a lifetime or across generations. Now that we know that it is entirely plausible to extract sensible prototypes from exemplars, we can start refining our approach (for psycholinguistic evaluation of the role property aggregates play in prototype formation and lexeme selection, see (Arppe et al. In preparation).

For example, the statistical method we have used requires us to provide, say, 250 annotated examples per structure, but human beings may well be able to generate a prototype from far fewer exemplars. In fact, in this paper, we may well have exaggerated the analytical challenge human beings face in reality: it is unlikely that humans are ever confronted with the task of learning all the data contained in our data frames at one single point in time (cf. Arppe and Baayen

2011). Work with developmental corpora, covering the stages from early childhood over adolescence to adulthood, once they become available would make it possible to take into account the incremental, time-dependent dimension of prototype formation, honouring the probably cumulative growth of a prototype out of exemplars over time. At the same time, work with substantially larger data sets that obviate the need for annotation and allow to model raw co-occurrence data without risking data sparsity would provide us with the opportunity to investigate the effect of incorporating in a model individual words as they appear in the sentences, instead of their analytical abstractions.

Finally, repeating this abstraction procedure for a range of phenomena from phonology, morphology and syntax and validating the categorization predictions experimentally would help us understand whether a particular level of abstraction would yield better categorization results for one specific phenomenon than for another. This would contribute to our understanding of the extent to which language really is best thought of as a single system, with the same cognitive operations active at all levels.

Acknowledgements

The authors would like to thank the audiences at the Slavic Cognitive Linguistics Conference 2009 in Prague (Czech Republic), the Linguistic Evidence 2010 Conference in Tübingen (Germany), the 7th International Conference on the Mental Lexicon in Windsor, Ontario (Canada), the Interdisciplinary Verb 2010 Workshop in Pisa (Italy), and the Linguistic Seminar Series Fall 2010 at Northumbria University (UK). We also thank Neil Bermel, Joan Bybee, Laura Janda, John Newman, the reviewers for Cognitive Linguistics – and the associate editor in particular – for valuable comments on earlier versions of this paper. The work of the second author was supported by a postdoctoral fellowship from the University of Helsinki as well as a grant for research collaboration abroad from the Academy of Finland (#136322/2009).

References

Apresjan, Jurij D., Ol'ga Ju. Boguslavskaja, Irina V. Levontina, Elena V. Uryson, Marina Ja. Glovinskaja & Tat'jana V. Krylova. 1999. *Новый объяснительный словарь синонимов русского языка.* [New explanatory dictionary of synonyms in Russian]*, Vol. 1.* Moskva: Škola "Jazyki Russkoj Kul'tury".

Arppe, Antti. 2002. The usage patterns and selectional preferences of synonyms in a morphologically rich language. In A. Morin & Pascale Sébillot (eds.), *JADT–2002: 6èmes*

journées internationales d'analyze statistique des données textuelles, Vol. 1, 21–32. Rennes: Institut National de Recherche en Informatique et en Automatique.

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, 44. Available at: <http://urn.fi/URN:ISBN:978-952-10-5175-3>.

Arppe, Antti. 2009. Linguistic choices vs. probabilities – how much and what can linguistic theory explain? In S. Featherston & S. Winkler (eds.), *The Fruits of Empirical Linguistics, Volume 1: Process*, 1–20. Berlin: de Gruyter.

Arppe, Antti. 2012. polytomous: Polytomous logistic regression for fixed and mixed effects. R package version 0.1.6.

Arppe, Antti. 2013. exemplars2prototypes: Vignette with instructions and R functions for extracting exemplary exemplars and prototypes. Available at: <http://cran.r-project.org/web/packages/polytomous/vignettes/exemplars2prototypes.pdf>.

Arppe, Antti & R. Harald Baayen. 2011. Statistical classification and principles of human learning. Paper presented at the International Conference on Quantitative Investigations in Theoretical Linguistics (QITL-4), Berlin, March 30, 2011.

Arppe, Antti & Dagmar Divjak. In preparation. Synonymy is both gradient and context-dependent.

Arppe, Antti, Ewa Dąbrowska & Dagmar Divjak. In preparation. Man versus machine. Using corpus-derived probabilities to predict lexical choice. (authors listed alphabetically)

Arppe, Antti, Petar Milin & R. Harald Baayen with contributions from Peter Hendrix. 2012. ndl: *Naive Discriminative Learning*. R package version 0.1.6. Available at: <http://CRAN.R-project.org/package=ndl>.

Baayen, R. Harald. 2007. Storage and computation in the mental lexicon. In G. Jarema and G. Libben (eds.), *The Mental Lexicon: Core Perspectives*, 81–104. Amsterdam: Elsevier.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data. A practical introduction to* statistics. Cambridge: Cambridge University Press.

Baayen, R. Harald. 2011. Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics* 11. 295–328.

Baayen, R. Harald, Petar Milin, Dušica Filipović-Đurđević, Peter Hendrix & Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118. 438–482.

Barsalou, Lawrence W. 1990. On the indistinguishability of exemplar memory and abstraction in category representation. In T. K. Srull & R. S. Wyer (eds.), *Advances in social cognition, Volume III: Content and process specificity in the effects of prior experiences*, 61–88. Hillsdale, NJ: Lawrence Erlbaum Associates.

Baetu, Irina & Thomas R. Shultz. 2010. Development of prototype abstraction and exemplar memorization. In S. Ohlsson & R. Catrambone (eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society,* 814–819. Austin, TX: Cognitive Science Society.

Bod, Rens. 2009. From Exemplar to Grammar: A Probabilistic Analogy-based Model of Language Learning. *Cognitive Science* 33(4). 752–793.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007. Predicting the Dative Alternation. In G. Boume, I. Kraemer & Joost Zwarts (eds.), *Cognitive Foundations of Interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bybee, Joan. 2010. *Language, Usage and Cognition.* Cambridge: Cambridge University Press.

Carey, Susan & Elsa Bartlett. 1978. Acquiring a single new word. *Papers and Reports on Child Language Development* 15. 17–29.

Cochran, William G. 1952. The χ2 Test of Goodness of Fit. *The Annals of Mathematical Statistics* 23(3), September. 315–345.
Cochran, William G. 1954. Some Methods for Strengthening the Common χ2 Tests. *Biometrics* 10(4), December. 417–451.
Cohen, Jacob, Patricia Cohen, Stephen G. West & Leona S. Aiken. 2003. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd edition). Mahwah: Lawrence Erlbaum Associates.
Connexor. 2007. List of morphological, surface-syntactic and functional syntactic features used in the linguistic analysis. [Web documentation] URL: <http://www.connexor.com/demo/doc/fifdg3-tags.html> [Accessed May 2007] and URL: <http://www.connexor.com/demo/doc/enfdg3-tags.html> [Accessed June 2007].
Dąbrowska, Ewa. 2008. The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: An empirical test of usage-based approaches to morphology. *Journal of Memory and Language* 58. 931–951.
Dennett, Daniel, C. 1991. *Consciousness explained*. Toronto: Little, Brown and Company.
Divjak, Dagmar. 2004. *Degrees of Verb Integration. Conceptualizing and Categorizing Events in Russian.* Unpublished PhD dissertation, K.U. Leuven (Belgium).
Divjak, Dagmar. 2010. *Structuring the lexicon: a clustered model for near-synonymy* (Cognitive Linguistics Research). Berlin: Mouton de Gruyter.
Divjak, Dagmar & Stefan Th. Gries. 2006. Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2(1). 23–60.
Divjak, Dagmar & Stefan Th. Gries. 2008. Clusters in the mind? Converging evidence from near synonymy in Russian. *The Mental Lexicon* 3(2). 188–213.
Divjak, Dagmar & Nick Fieller. Forthcoming. Finding structure in linguistic data. In Justyna Robinson and Dylan Glynn (eds), *Empirical Approaches to Polysemy and Synonymy* (Human Cognitive Processes Series). Amsterdam: Benjamins.
Ellis, Nick C. & Fernando Ferreira-Junior. 2009a. Construction learning as a function of frequency, frequency distribution, and function. *Modern Language Journal* 93. 370–385.
Ellis, Nick C. & Fernando Ferreira-Junior. 2009b. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7. 188–221.
Estes, William K. 1994. *Classification and Cognition*. New York/Oxford: Oxford University Press.
Goldberg, Adele. 2006. *Constructions at work. The nature of generalization in language.* Oxford: Oxford University Press.
Gries, Stefan Th. 2003a. *Multifactorial analysis in corpus linguistics: a study of Particle Placement.* London and New York: Continuum Press.
Gries, Stefan Th. 2003b. Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics* 1. 1–27.
Gries, Stefan Th. 2009. *Statistics for linguistics with R: a practical introduction*. Berlin and New York: Mouton de Gruyter.
Griffiths, Thomas L., Kevin R. Canini, Adam N. Sanborn & Daniel J. Navarro. 2007. Unifying rational models of categorization via the hierarchical Dirichlet process. In D. S. McNamara & J. G. Trafton (eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, 323–328. Mahwah, NJ: Erlbaum.
Grondelaers, Stefan, Dirk Speelman & Dirk Geeraerts. 2002. Regressing on *er*. Statistical analysis of text and linguistic variation. In A. Morin and P. Sébillot (eds.), *JADT-2002: 6èmes journées internationales d'analyze statistique des données textuelles*,

Vol. 1, 335–346. Rennes: Institut National de Recherche en Informatique et en Automatique.
Harrell, Frank E. 2001. *Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression and Survival Analysis*. New York, NY: Springer-Verlag.
Harrington, Jonathan. 2006. An acoustic analysis of 'happy-tensing' in the Queen's Christmas broadcasts. *Journal of Phonetics* 34. 439–457.
Hawkins, Jeff & Sandra Blakeslee. 2004. *On intelligence*. New York: Henry Holt and Company.
Hintzman, Douglas, L. 1986. "Schema abstraction" in a multiple-trace memory model. *Psychological Review* 93(4). 411–428.
Hosmer, David W., Jr. & Stanley Lemeshow. 2000. *Applied Regression Analysis* (2nd edition). New York, NY: Wiley.
Inkpen, Diana. 2004. *Building a Lexical Knowledge-Base of Near-Synonym Differences*. PhD dissertation, Department of Computer Science, University of Toronto.
Inkpen, Diana & Graeme Hirst. 2006. Building and Using a Lexical Knowledge-Base of Near-Synonym Differences. *Computational Linguistics* 32(2). 223–262.
Järvinen, Timo & Pasi Tapanainen. 1997. *A Dependency Parser for English*. TR-1, Technical Reports of the Department of General Linguistics, University of Helsinki, Finland.
Johnson, Keith. 2008. Quantitative methods in linguistics. Malden/Oxford/Victoria: Wiley-Blackwell.
Katz, Jerrold J. & Paul M. Postal. 1964. *An integrated theory of linguistic descriptions*. Cambridge, MA: MIT Press.
Labov, William. 1973. The boundaries of words and their meanings. In C. Bailey and R. Shuy (eds.), New Ways of Analyzing Variation in English. Washington, DC: Georgetown University Press, 340–373. Reprinted in B. Aarts et al. (eds.), Fuzzy Grammar. Oxford: Oxford University Press, 67–90.
Lakoff, George. 1987. *Women, fire, and dangerous things: what categories reveal about the mind.* Chicago, IL: University of Chicago Press.
Landauer, Thomas & Susan T. Dumais. 1997. A solution to Plato's problem: the Latent Semantic Analysis theory for acquisition, induction and representation of knowledge. *Psychological Review* 104(2). 211–240.
Langacker, Ronald. 1987. *Foundations of Cognitive Grammar: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.
Langacker, Ronald. 2010. How Not to Disagree: The Emergence of Structure from Usage. In K. Boye & E. Engberg-Pedersen (eds.), *Language Usage and Language Structure* (Trends in Linguistics Studies and Monographs 213), 107–143. Berlin and New York: De Gruyter Mouton.
Love, Bradley C., Douglas L. Medin & Todd M. Gureckis. 2004. SUSTAIN: A network model of category learning. *Psychological Review* 111. 309–332.
Lyons, John. 1977. *Semantics*. Cambridge: Cambridge University Press.
Medin Douglas L. & Marguerite M. Schaffer. 1978. Context theory of classification learning. *Psychological Review* 85. 207–238.
Menard, Scott. 1995. *Applied Logistic Regression Analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences 07–106. Thousand Oaks: Sage Publications.
Murphy, Gregory L. 2002. *The big book of concepts*. Cambridge, MA: MIT Press.
Nosofsky, Robert. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology*, 115(1). 39–37.

Nosofsky, Robert. 1992. Exemplars, prototypes and similarity rules. In A. F. Healy, S. M. Kosslyn & R. M. Shiffrin (eds.), *Essays in honor of William K. Estes: Vol I. From learning theory to connectionist theory,* 149–167. Hillsdale, NJ: Erlbaum.
O'Brien, Robert M. 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality and Quantity*. 41(5). 673–690.
Ožegov, Sergej I. & Natal'ja Ju. Švedova. 1999. *Толковый словарь русского языка* [Explanatory dictionary of the Russian language]. Moskva: Azbukovnik.
Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee and P. Hopper (eds.), *Frequency effects and the emergence of lexical structure,* 137–157. John Benjamins, Amsterdam.
Pulman, Stephen G. 1983. *Word Meaning and Belief.* London: Croom Helm.
Rifkin, Ryan & Aldebaro Klautau. 2004. In Defense of One-Vs-All Classification. *Journal of Machine Learning Research* 5. 101–141.
Rosch, Eleanor. 1973. Natural Categories. *Cognitive Psychology* 4. 328–350.
Rosseel, Yves. 2002. Mixture models of categorization. *Journal of Mathematical Psychology*, 46. 178–210.
Saffran, Jenny, R., Richard N. Aslin, & Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274. 1926–1928.
Sankoff, David. 1978. Probability and Linguistic Variation. *Synthèse* 37. 217–238.
Taylor, John R. 1989. *Linguistic Categorization: Prototypes in Linguistic Theory. Second Edition*. Oxford: Clarendon Press.
Tapanainen, Pasi & Timo Järvinen. 1997. A non-projective dependency parser. *Proceedings of the 5th Conference on Applied Natural Language Processing*. Association of Computational Linguistics. 64–71.
Theil, Henri. 1970. On the Estimation of Relationships Involving Qualitative Variables. *The American Journal of Sociology* 76(1). 103–154.
Tuggy, David. 2007. Schematicity. In D. Geeraerts & H. Cuyckens (eds), *The Oxford Handbook of Cognitive Linguistics*, 82–116. Oxford: OUP.
Vanpaemel, Wolf & Gerrit Storms. 2008. In search of abstraction: the varying abstraction model of categorization. *Psychonomic Bulletin and Review* 15(4). 732–749.
Vanpaemel, Wolf & Gerrit Storms. 2010. Abstraction and model evaluation in category learning. *Behavioral Research Methods* 42(2). 421–437.
Verbeemen, Timothy, Wolf Vanpaemel, Sven Pattyn, Gerrit Storms & Tom Verguts. 2007. Beyond exemplars and prototypes as memory representations of natural concepts: a clustering approach. *Journal of Memory and Language* 56. 537–554.
Verhagen, Arie. 2005. *Constructions of Intersubjectivity*. Oxford: OUP.
Wittgenstein, Ludwig. 1953/2001. *Philosophical Investigations*. Blackwell Publishing.

Corpora

AC Amsterdam Corpus, compiled by A. Barentsen. Available on CD from the compiler on request.
RNC Russian national corpus. Available on-line at URL <http://www.ruscorpora.ru/en/>.
Helsingin Sanomat. 1995. ~22 million words of Finnish newspaper articles published in Helsingin Sanomat during January–December 1995. Compiled by the Research Institute for

the Languages of Finland [KOTUS] and CSC – IT Center for Science, Finland. Available on-line at URL: <http://www.csc.fi/kielipankki/>.

SFNET. 2002–2003. ~100 million words of Finnish internet newsgroup discussion posted during October 2002–April 2003. Compiled by Tuuli Tuominen and Panu Kalliokoski, Computing Centre, University of Helsinki, and Antti Arppe, Department of General Linguistics, University of Helsinki, and CSC – IT Center for Science, Finland. Available on-line at URL: <http://www.csc.fi/kielipankki/>.