Extracting prototypes from corpus data: a distributional account of representing near-synonymous verbs

Dagmar Divjak
University of Sheffield
Sheffield, United Kingdom
d.divjak@sheffield.ac.uk

Antti Arppe
University of Helsinki
Helsinki, Finland
antti.arppe@helsinki.fi

Abstract

In cognitive linguistics, *prototype* theory is currently one of the dominant views of how linguistic categories are stored and represented as cognitive structures in the brain. Yet two problems arise: Cognitive linguistics is a usage-based theory but has thus far not attempted to show how prototypes can be observed in usage in a systematic way. Furthermore, the bulk of the research done has focused on prototypes for nouns, denoting tangible objects, rather than verbs that denote intangible events. In this paper, we simulate how abstract prototypes for verbs could be formed using statistical learning mechanisms that track frequency distributions in input on the basis of actual usage as observed in corpus data.

1 Introduction

Nearly four decades ago Eleanor Rosch (1973 and later work) demonstrated the inadequacy of necessary and sufficient attributes for item classification. Instead, she presented a prototype approach to categorization, a probabilistic feature approach with instances displaying different degrees of representativity and similarity to a 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 cognitive linguistics, *prototype* theory is one of the dominant views of how linguistic categories are stored and represented as cognitive structures in the brain (Taylor 1995). Yet, although cognitive linguistics actively promotes itself as a usage-based theory, thus far it has not been shown how prototypes can be observed in actual usage in a systematic and cognitively realistic way.

Moreover, the bulk of experimental and linguistic research done on prototype categorization has concentrated on nouns (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 neither stable in nor independent of time. In addition, nouns typically denote tangible objects, whereas verbs name intangible events. And thirdly, verbs render relational concepts, which implies that they are more susceptible to their 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 procedure of extracting verbal prototypes from language use.

2 Methodology

We do so by statistically modeling large annotated datasets of exemplars and gradually reducing exemplars while abstracting properties. To this end, we build upon the results of the application of a multivariate statistical technique, *polytomous logistic regression* (see e.g. Arppe 2008) according to the *one-vs-rest* heuristic (Rifkin & Klautau 2004) which was used to study the contextual similarities and differences of two sets of Russian and Finnish near-synonyms expressing TRY and THINK. These two synonym sets from two typologically distinct languages have been selected for the practical reason that they have been the object of recent large-scale corpus-based studies (Arppe 2008 and Divjak 2010) exploring the phenomenon near-synonymy from different angles, which have produced extensive

datasets for further analyses such as the one presented here.

2.1 Data

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 the Amsterdam Russian Corpus, the Russian National Corpus and (selected) Internet pages. In all, there were 1,351 occurrences of this syntactically homogenous category (i.e. all verbs share the same argument structure). Depending on the frequency of the verb, between 100 and 250 examples were annotated per verb.

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 (Helsingin Sanomat 1995) and six months of Internet newsgroup discussion (SFNET 2002-2003). 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 from Divjak & Gries (2006). This scheme captures all information provided at the sentence level by tagging for morphological properties of the finite verb and the infinitive (tense, aspect, mode), syntactic properties of the sentences (sentence type, clause type) and semantic properties of the infinitive (semantic type of subject, properties of the event denoted by the infinitive, controllability of the infinitive action) as well as optional elements (adverbs, particles, negation). The final tagset contains 14 variables amounting to 87 variable categories. This annotation scheme thus contains all elements ecountered within sentence boundaries and can, as such, be transferred to the annotation of other verbs, e.g. verbs expressing INTENTION (Divjak 2006, 2010) or RESULT (Divjak 2003, 2010).

For Finnish, the 3,404 examples were first morphologically and syntactically analyzed using an implementation of the Functional-Dependency Grammar (FDG) parser (Tapanainen & Järvinen 1997) for Finnish, after which 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. The morphological level of analysis of the node verb covered subtypes of infinitive and participle, non-finite case, number and possessive suffix (indicating person and number), polarity, voice, mood, simplex tense, and finite person-number, whereas that of the entire verb chain of which the THINK verb was part of concerned polarity, voice, mood, an aggregate of person and number marking for both finite or non-finite verb forms, and surface-syntactic role. The syntactic argument types follow those of the FDG formalism, and the semantic and structural subtyping was a combination of various schemes including WordNet (Miller et al. 1990), several prior Finnish studies (Pajunen 2001, Kangasniemi 1992 and Flint 1980) and an evidence-based bottom-up classification procedure suggested by Hanks (1996).

Although the two analysis schemes have different starting points (i.e. an argument structurally homogenous 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 in its entirety. Moreover, using such two distinct schemes is a test of the overall robustness of the statistical modelling and analysis, provided we are able to produce effectively similar results.

2.3 Statistical modeling

We modeled the annotated corpus data using polytomous logistic regression (see e.g. Arppe 2008).[1] The one-vs-rest heuristic (Rifkin & Klautau 2004) distinguishes each member of the set without requiring a baseline category and directly provides lexeme-specific odds with respect to selected variables (representing linguistic properties). It models probabilities of occurrence given a particular context. The variable parameters it estimates can be naturally interpreted as *odds* (Harrell 2001). As a simple selection rule, the verb receiving the highest estimated probability

[1] Since *multinomial logistic regression* is often used to refer in effect to only a particular heuristic out of many possible ones, i.e. where a set of (binary) baseline models are fitted simultaneously and in relation to each other with a given algorithm, we use the term *polytomous logistic regression* modeling as an umbrella concept for any heuristic tackling polytomous (i.e. more than two alternatives) outcomes as long as it is based on logistic regression analysis, regardless of how the polytomous setting is broken down into a set of binary models and whether these component binary models are separately or simultaneously fitted (for an overview, see Arppe 2008).

is picked for any given context representing a cluster of properties, i.e., $arg\text{-}max_{Verb}[P(Verb|Context)]$. 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. Moreover, since the constituent binary logistic regression models are fit separately with the one-vs-rest heuristic, the sums of their instance-wise probabilities are not always exactly $\sum P=1.0$. Therefore, the verb-specific probabilities for each instance in both data sets are adjusted to satisfy this condition by dividing, instance-wise, each original lexeme-specific probability estimate by the sum of these estimates for that particular instance.

The original variable sets were pruned since the number of variables allowed in logistic regression is maximally 1/10 of the frequency of the rarest outcome. The selection strategy we adopted for the Russian TRY lexemes was to retain variables with a broad dispersion among the verbs. We required the overall frequency of the variable in the data to be at least 45 and to occur at least twice with all verbs. Additional technical restrictions excluded one variable for each fully complementary case (e.g. the aspect of the verb form) as well as variables with mutual pairwise association statistic Uncertainty Co-Efficient (Theil 1970) UC>0.5 (i.e. knowledge of one variable decreases more than ½ of the uncertainty concerning the other). In the end, 18 property variables remained.

For the Finnish THINK lexemes, a minimum overall frequency was required, in this case set at $n\geq24$. Pair-wise associations of individual properties were likewise carefully evaluated using UC, but due to the heterogeneity of the argument structure of the Finnish THINK verbs, occurrence with all four verbs was not required. Semantic subtypes were included only for the most frequent syntactic argument types, and many contextual property variables were lumped together, whenever possible and appropriate. In the end, 46 linguistic property variables were chosen for the full model, of which 10 were morphological, concerning the entire verb chain, 10 simple syntactic arguments (without any semantic subtypes), 20 combinations of syntactic arguments with semantic and structural subclassifications, and 6 semantic characterizations of the entire verb chains.

2.4 Model performance

In the case of the six Russian TRY verbs, 51.7% of all cases were correctly predicted (i.e. Recall) according to the prediction rule of selecting the verb with the highest estimated probability. The Recall rate for the four Finnish THINK verbs was 64.6%. Comparing these percentages to the 52.7% correct answers the average non-English US college applicant provided in a 4-way choice between semantically related verbs such as *imposed, believed, requested and correlated* (Landauer and Dumais 1997) confirms that the statistical models perform at a rate comparable to that of human beings.

3 Results

3.1 Property-wise verb-specific odds

The one-vs-rest analysis technique has two key attractive characteristics as stepping stones towards showing how prototype formation may be achieved on the basis of usage data.

Firstly, a model created with polytomous logistic regression provides probability estimates for the (proportional) occurrence of an outcome, such as a verb within some synonym set, given the contextual occurrence of some combination of linguistic properties incorporated in the model. Secondly, and crucially, 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*. Individual odds (parameter values) which are greater than 1.0 for some property and the singled-out verb can be interpreted to reflect the increased chances of occurrence of this verb when the property in question is present in the context. Conversely, odds less than 1.0 denote a decreased chance of the occurrence of this verb in such a context. As an example case, take the Russian *probovat'*, for which the property-wise odds are shown in Table 1.

PROPERTY/VERB	ODDS
(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)
INF….ASPECT_IMPERFECTIVE	**6.1:1**
INF….CONTROL_HIGH	(1:1.2)
INF….SEM_COMMUNICATION	**2.1:1**
INF….SEM_EXCHANGE	(1.4:1)
INF….SEM_METAPH…_MOTION	(1.5:1)
INF….SEM_METAPH…_PHYSICAL_EXCHANGE	(1:1.3)
INF….SEM_METAPH…_PHYSICAL_OTHER	(1.3:1)
INF….SEM_MOTION	(1.7:1)
INF….SEM_MOTION_OTHER	(2.6:1)
INF….SEM_PHYSICAL	**3.9:1**
INF….SEM_PHYSICAL_OTHER	**2.5:1**
SENTENCE.DECLARATIVE	1:2.8
SUBJECT.SEM_ANIMATE_HUMAN	(1.5:1)

Table 1: Odds for/against Russian *probovat'* (Odds in parentheses are non-significant)

3.2 Aggregating properties as a prototype

In the model, those verb-specific linguistic properties – *per definition* abstract generalizations – which have significant odds in favor of a verb 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 either near-synonym set.

For the Russian TRY verbs, out of a total of 1,351 individual property combinations, 660 combinations of a distinct verb plus a context type can be distinguished (reducing to 296 if the outcome verb is ignored), leading to 20 permissible property combinations with significantly favorable odds, and ultimately to as few aggregates of properties with such strongly favorable odds as there are verbs. For *probovat'*, the set of such properties with significant odds in favor of this verb occurring when they are evident in the context are boldfaced in Table 1. Note that only one of the three semantic characterizations of the infinitive can possibly be observed at the same time. Thus, the aggregate of properties in fact represents three permissible property combinations.

For the Finnish THINK verbs, out of a total of 3,404 individual property combinations, 2,196 combinations of distinct property clusters with (one of the) verbs can be identified, which reduces only slightly to 1,908 if the outcome verb is ignored. This is a result of the heterogeneity of the allowed argument structures of the Finnish THINK verbs (versus the syntactic homogeneity of the Russian TRY verbs), as well as the greater overall number of properties included in the analysis. Due to this syntactic heterogeneity and optionality of many arguments and properties, practically only a lower bound can be estimated of altogether at least 51 permissible combinations of properties with significant favorable odds for the four THINK lexemes, distributed as follows: *ajatella* (32), *miettiä* (8), *pohtia* (10), and *harkita* (1).

4 Discussion

The aggregated properties with significant odds in favor of a verb are, as a whole, manifestations of (the core of) a prototype for a verb. It is plausible to interpret the above properties for *probovat'* as conveying the notion of telling someone to *try* (using the perfective aspect hence signaling 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), and why it is the most frequently used TRY verb in mother-child interaction (Stoll corpus, see Divjak & Gries 2006).

This definition has been distilled from the extracted estimated odds over properties that predict which of the near-synonymous alternatives is most likely to be selected given a specific linguistic context. Over the past decade, numerous studies have been published supporting the claim that infants are equipped with powerful statistical language learning mechanisms. If speakers model input statistically, as is assumed by statistical learning (cf. Saffran et al. 1996), they may be operating with similar prototypes as the regression technique outputs.

Nevertheless, a caveat needs to be expressed. We have aimed to model produced language systematically by means of a statistical heuristic, regression analysis, yet the heuristic by which this model is constructed and the constituent binary logistic regression models and mathematical algorithms by which they are optimized to fit to the data were not designed to mimic cognitive behavior. The resulting model fits descriptions that linguists feel are appropriate for the data, but the underlying mechanics of regression analysis lacks cognitive grounding other than the fact that human beings seem able to detect statistical regularities in input.

References

Apresjan, J. D. 1999, *Новый объяснительный словарь синонимов русского языка. Vol. I.* Moskva: Škola "Jazyki Russkoj Kul'tury".

Arppe, A. 2008. *Univariate, bivariate, and multivariate methods in corpus-based lexicography – A study of synonymy*. Publications of the Dep't of General Linguistics, University of Helsinki, 44.

Divjak, D. 2003. Слова о словах. К вопросу о неточных синонимах «умудриться», «ухитриться», «исхитриться» и «изловчиться. [Words on Words. The Case of Four Near Synonyms] In: Soldatjenkova, T., Waegemans, E. For East is East. Liber Amicorum Wojciech Skalmowski. Leuven-Paris: Peeters, 345-365. [Orientalia Lovaniensia Analecta].

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

Divjak, D. 2010 (in press). *Structuring the Lexicon. A Clustered Model for Near-Synonymy.* Berlin/New York: Mouton de Gruyter. Cognitive Linguistics Research [CLR] 43.

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

Flint, A. 1980. *Semantic Structure in the Finnish Lexicon: Verbs of Possibility and Sufficiency*. Helsinki: Suomalaisen Kirjallisuuden Seura (SKST 360).

Hanks, P. 1996. Contextual Dependency and Lexical Sets. *International Journal of Corpus Linguistics*, 1:1, 75-98.

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

Kangasniemi, H. 1992. *Modal Expressions in Finnish*. Studia Fennica, Linguistica 2. Helsinki: Suomalaisen Kirjallisuuden Seura.

Landauer, T. & S. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psych. Review*, 104 (2): 211–240.

Miller, G. A. 1990. Nouns in WordNet: a lexical inheritance system. (revised August 1993). *International Journal of Lexicography*, 3:4, 245–264.

Pajunen, A. 2001. *Argumenttirakenne: Asiaintilojen luokitus ja verbien käyttäytyminen suomen kielessä*. Suomi 187. Helsinki: Suomalaisen Kirjallisuuden Seura.

Pulman, S. G. 1983. *Word Meaning and Belief.* London: Croom Helm.

Rifkin, R. & A. Klautau. 2004. In Defense of One-Vs-All Classification. *Journal of Machine Learning Research*, 101-141.

Rosch, E. 1973. Natural Categories. *Cognitive Psychology* 4: 328–350.

Saffran, Jenny R, Richard N. Aslin and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274: 1926-1928.

Tapanainen, P. & T. Järvinen. 1997. A non-projective dependency parser. *Proceedings of the 5th Conference on Applied Natural Language Processing*. Assoc. of Computational Linguistics, 64–71.

Taylor, J. R. 1995. *Linguistic Categorization: Prototypes in Linguistic Theory. 2nd Edition.* Oxford: Clarendon Press.

Theil, Henri 1970. On the Estimation of Relationships Involving Qualitative Variables. *The American Journal of Sociology*, 76:1, 03-154.