Linguistic choices vs. probabilities – how much and what can linguistic theory explain?

Antti Arppe
Department of General Linguistics
University of Helsinki
antti.arppe@helsinki.fi

1. Introduction and background

A question of general theoretical interest in linguistics is what is the relationship between naturally produced language, evident in e.g. corpora, and the posited underlying language system that governs such usage. This concerns on the one hand the use and choice among lexical and structural alternatives in language, and on the other the underlying explanatory factors, following some theory representing language as a cohesive system. A subsequent subservient methodological challenge is how this can be modeled using appropriate statistical methods. The associated question of general theoretical import is to what extent we can describe the observed usage and the variation it contains in terms of the selected analytical features that conventional linguistic theory incorporates and works upon. The practical purpose of this paper is to present a case study elucidating how multivariate statistical models can be interpreted to shed light on these questions, focusing on a set of near-synonyms as the particular type of linguistic alternation. With multivariate modeling, I mean two distinct things. Firstly, I imply the use of multiple linguistic variables from a range of analytical levels and categories, instead of only one or two, in order to study and explain some linguistic phenomenon. Secondly, I mean with this term the use of multivariate statistical methods such as polytomous logistic regression. In the following introduction, I will first present research demonstrating that one and the same linguistic phenomenon can be associated with, and appear to be explainable in terms of a wide range of different variables from various levels of linguistic analysis. Next, I will note research indicating that satisfactory explanations of such linguistic phenomena requires multivariate (multicausal) models, i.e. the incorporation of all of these variables at the same time in the analysis. This leads us to the final and central question of how much of the phenomena we can in the end account for with the fullest set of explanatory variables available to us in current linguistic analysis.

In the modeling of lexical choice among semantically similar words, specifically near-synonyms, it has been suggested in computational theory that (at least) three levels of representation would be necessary to account for fine-grained meaning differences and the associated usage preferences (Edmonds and Hirst 2002: 117-124). These are 1) a conceptual-semantic level, 2) a subconceptual/stylistic-semantic level, and 3) a syntactic-semantic level, each corresponding to increasingly more detailed representations, i.e. granularity, of (word) meaning. The last, syntactic-semantic level (3) in such a *clustered model of lexical knowledge* concerns the combinatorial preferences of individual words in forming written sentences and spoken utterances. At this level, it has been shown in (mainly) lexicographically motivated corpus-based studies of actual lexical usage that semantically similar words differ significantly as to the different types of context in which are used. This has been observed to concern 1) lexical context (e.g. English adjectives *powerful* vs. *strong* in Church et al. 1991), 2)

syntactic argument patterns (e.g. English verbs *begin* vs. *start* in Biber, Conrad and Reppen 1998: 95-100), and 3) the semantic classification of some particular argument (e.g. the subjects/agents of English *shake/quake* verbs in Atkins and Levin 1995), as well as 4) the rather style-associated text types or registers (e.g. English adjectives *big* vs. *large* vs. *great* in Biber, Conrad and Reppen 1998: 43-54). In addition to these studies that have focused on English, with its minimal morphology, it has also been shown for languages with extensive morphology, such as Finnish, that similar differentiation is evident as to 5) the inflectional forms and the associated morphosyntactic features in which synonyms are used (e.g. the Finnish adjectives *tärkeä* and *keskeinen* 'important, central' in Jantunen 2001; and the Finnish verbs *miettiä* and *pohtia* 'think, ponder, reflect, consider' in Arppe 2002, Arppe and Järvikivi 2007). Recently, in their studies of Russian near-synonymous verbs denoting TRY as well as INTEND, Divjak (2006) and Divjak and Gries (2006) have shown that there is often more than one type of these factors in play at the same time. Divjak and Gries' subsequent conclusion is that it is necessary to observe all categories together and in unison rather than separately one by one.

Similar corpus-based work has also been conducted on the syntactic level concerning *constructional alternations* (referred alternatively to as *synonymous structural variants* in Biber, Conrad and Reppen 1998: 76-83), often from starting points which would be considered to be anchored more within theoretical linguistics. Constructional alternations resemble lexical synonymy in that the essential associated meaning is understood to remain for the most part constant regardless of which of the alternative constructions is selected, though they may differ with respect to e.g. some pragmatic aspect such as focus. Relevant studies concerning these phenomena have been conducted by e.g. Gries (2003a) concerning the English verb-particle placement, i.e. [$V\ P\ NP_{DIRECT\_OBJECT}$] vs. [$V\ NP_{DIRECT\_OBJECT}\ P$], and Gries (2003b) as well as Bresnan et al. (2007) concerning the English dative alternation, i.e. [$GIVE\ NP_{DIRECT\_OBJECT}\ PP_{INDIRECT\_OBJECT}$] vs. [$GIVE\ NP_{INDIRECT\_OBJECT}\ NP_{DIRECT\_OBJECT}$], to name but just a few.

With the exception of Gries (2003a, 2003b), Bresnan et al. (2007), Divjak (2006), and Divjak and Gries (2006), the aforementioned studies have in practice been monocausal, focusing on only one linguistic category or even a singular feature within a category at a time in the linguistic analysis applied. Though Jantunen (2001, 2004) does set out to cover a broad range of feature categories and notes that a linguistic trait may be evident at several different levels of context at the same time (2004: 150-151), he does not quantitatively evaluate their interactions. Bresnan et al. (2007) have suggested that such a tendency for reductive theories would result from pervasive correlations among the possible explanatory variables in the available data. Indeed, Gries (2003a: 32-36) has criticized this traditional tendency for monocausal explanations and demonstrated convincingly that such individual univariate analyses are insufficient and often even mutually contradictory. As a necessary remedy in order to attain scientific validity in explaining the observed linguistic phenomena, he has argued forcefully for a holistic approach using multifactorial setups covering a representative range of linguistic categories, leading to and requiring the exploitation of multivariate statistical methods. In such an approach, linguistic choices, whether synonyms or alternative constructions, are understood to be determined by a *plurality* of factors, in *interaction* with each other.

Furthermore, as has been pointed out by Divjak and Gries (2006), the majority of the above and other synonym studies appear to focus on word pairs, perhaps due to the methodological simplicity of such setups. The same criticism of limited scope applies also to studies of constructional alternations, including e.g. Gries' (2003a) own study on English particle placement. However, it is clearly evident in lexicographical descriptions such as dictionaries that there are often more than just two members to a synonym group, and this supported by experimental evidence (Divjak and Gries, forthcoming). Likewise, it is quite easy to come up with examples of constructional alternations with more than two conceivable and fully possible variants, e.g. in word order. This clearly motivates a shift of focus in synonym studies from word pairs to sets of similar lexemes with more than two members; the same naturally applies also to the study of constructional alternations.

Finally, Bresnan (2007) has suggested that the selections of alternatives in a context, i.e. lexical or structural outcomes for some combinations of variables, are generally speaking probabilistic, even though the individual choices in isolation are discrete. In other words, the workings of a linguistic system, represented by the range of variables according to some theory, and its resultant usage would in practice not be categorical, following from exception-less rules, but rather exhibit degrees of potential variation which becomes evident over longer stretches of linguistic usage. This is manifested in the observed proportions of occurrence among the possible alternating structures, given a set of contextual features. Bresnan (2007) uses logistic regression to model and represent these proportions as estimated expected probabilities, producing a continuum of variation between the practically categorical extremes (see Figure 1 in Bresnan 2007: 77, based on results from Bresnan et al. 2007). Moreover, both Gries (2003b) and Bresnan (2007) have shown that there is evidence for such probabilistic character both in natural language use in corpora as well as in language judgements in experiments, and that these two sources of evidence are convergent.

Nevertheless, one may question whether Bresnan's (2007) results entail that an idealization of linguistic system as a whole, as knowledge incorporated in an *ideally complete* theoretical model that describes its workings in its entirety (disregarding if such is in practice attainable at all), with syntax as one constituent level interacting with phonological, prosodic, lexical, semantic, pragmatic and extralinguistic ones, need be fundamentally non-categorical (see e.g. Yang, 2008, and the references therein). In any case, how linguistic probabilities are represented within speakers' minds, how they come about as either individual linguistic judgments, or as proportions in language usage , and how they (inevitably) change over time within a linguistic community, is beyond the scope of this paper.

However, the aforementioned studies by Bresnan and Gries, too, have concerned only dichotomous outcome alternatives. Consequently, my intention is to extend this line of research to a polytomous setting involving the lexical choice among more than two alternatives, using as a case example the most frequent near-synonymous THINK lexemes in Finnish, namely *ajatella*, *miettiä*, *pohtia*, and *harkita* 'think, reflect, ponder, consider'. Furthermore, in line with the aforementioned previous research, I will include in the analysis a broad range of contextual features as explanatory variables. Thus, I will cover 1) the morphological features of both the selected verbs and the verb-chains they may be part of, 2) the entire syntactic argument structure of the verbs, 3) the semantic subclassifications of the individual argument types, 4) the

semantic characterizations of the entire verb-chains in which the verbs occur, as well as 5) extralinguistic features such as medium.

2. Research corpus as well as linguistic and statistical analysis methods

As my research corpus, I selected two months worth (January–February 1995) of written text from Helsingin Sanomat (1995), Finland's major daily newspaper, and six months worth (October 2002 – April 2003) of written discussion in the SFNET (2002-2003) Internet discussion forum, namely regarding (personal) relationships (`sfnet.keskustelu.ihmissuhteet`) and politics (`sfnet.keskustelu.politiikka`). The newspaper subcorpus consisted altogether of 3,304,512 words of body text, excluding headers and captions (as well as punctuation tokens), and included 1,750 representatives of the selected THINK verbs. In turn, the Internet subcorpus comprised altogether 1,174,693 words of body text, excluding quotes of previous postings as well as punctuation tokens, adding up to 1,654 representatives of the selected THINK verbs. The individual overall frequencies among the THINK lexemes in the research corpus were 1492 for *ajatella*, 812 for *miettiä*, 713 for *pohtia*, and 387 for *harkita*.

The details of the various stages and levels of linguistic analysis applied to this research corpus are covered at length in Arppe (2008), but I will briefly cover the main points also here. The research corpus was first automatically morphologically and syntactically analyzed using a computational implementation of Functional Dependency Grammar (Tapanainen and Järvinen, 1997, Järvinen and Tapanainen 1997) for Finnish, namely the FI-FDG parser (Connexor 2007). After this automatic analysis, all the instances of the THINK lexemes together with their syntactic arguments were manually validated and corrected, if necessary, and subsequently supplemented with semantic classifications by hand. Each nominal argument (in practice nouns or pronouns) was semantically classified into one of the 25 top-level *unique beginners* for (originally English) nouns in WordNet (Miller 1990). Furthermore, subordinate clauses or other phrasal structures assigned to the PATIENT argument slot were classified following Pajunen (2001) into the traditional types of either participles, infinitives, indirect questions, clause propositions indicated with the subordinate conjunction *että* 'that', or direct quotes with attributions of the speaker using one of the THINK lexemes (e.g. *"..." miettii/pohtii joku* '"..." thinks/ponders somebody'). This covered satisfactorily AGENTs, PATIENTs, SOURCEs, GOALs and LOCATIONs among the frequent syntactic argument types as well as INSTRUMENTs and VOCATIVEs among the less frequent ones.

However, other syntactic argument types which were also frequent in the context of the THINK lexemes, indicating MANNER, TIME (as a moment or period), DURATION, FREQUENCY and QUANTITY, had a high proportion of adverbs, prepositional/postpositional phrases and subordinate clauses (or their equivalents based on non-finite verb forms). These argument types were semantically classified following the *ad hoc* evidence-driven procedure proposed by Hanks (1996), in which one scrutinizes and groups the individual observed argument lexemes or phrases in a piece-meal fashion. In Hanks' approach, as contextual examples accumulate, one generalizes semantic classes out of them, possibly reanalyzing the emergent classification if need be, without attempting to apply some prior theoretical model.

Only in the case of MANNER arguments did several levels of granularity emerge at this stage in the semantic analysis. Even though clause-adverbials (i.e. META-comments such as *myös* 'also', *kuitenkin* 'nevertheless/however' and *ehkä* 'maybe' as well as subordinate clauses initiated with *mutta* 'but' and *vaikka* 'although') were also relatively quite frequent as an argument type, they were excluded from this level of analysis due to their generally parenthetical nature.

Furthermore, as an extension to Arppe (2006) the verb chains which the THINK lexemes form part of were semantically classified with respect to their modality and other related characteristics, following Kangasniemi (1992) and Flint (1980). Likewise, those other verbs which are syntactically in a co-ordinated (and similar) position in relation to the THINK lexemes were also semantically classified, following Pajunen (2001). Moreover, with respect to morphological variables, I chose to supplement analytic features characterizing the entire verb chain of which the THINK lexemes were components of, concerning polarity (i.e. AFFIRMATION in addition to the explicitly marked NEGATION), voice, mood, tense and person/number. Thus, if a non-finite form of the THINK lexemes is an integral part of a verb-chain, which contains constituents that are explicitly marked with respect to person-number or any of the other features normally associated only with finite verb forms, such features will be considered to apply for the non-finite THINK form, too. In addition, the six distinct person/number features (e.g. FIRST PERSON SINGULAR, FIRST PERSON PLURAL, SECOND PERSON SINGULAR, and so on) were decomposed as a matrix of three person features (FIRST vs. SECOND vs. THIRD) and two number features (SINGULAR vs. PLURAL). A representative overview of the entire range of feature categories and individual features applied in the linguistic analysis of the research corpus is presented in Table 1.

Table 1. Overview of the various contextual feature categories and individual features included in the linguistic analysis of the research corpus; features in (parentheses) have been excluded from some models in the multivariate statistical analyses due to their low frequency, e.g. POTENTIAL mood, or high level of association with some other feature, e.g. IMPERATIVE mood (in comparison with SECOND person) on the verb-chain general level, or a complementary or near-complementary distribution, e.g. AFFIRMATION (vs. NEGATION), also on the verb-chain general level; furthermore, some features in {brackets} have been lumped together in some models, e.g. human INDIVIDUALs and GROUPs under syntactic AGENTs are sometimes collapsed together as HUMAN referents.

Node-specific morphological features	
infinitive subtype	FIRST INFINITIVE (*-A*), SECOND INFINITIVE (*-E-*), THIRD INFINITIVE (*-MA-*), FOURTH INFINITIVE (*-minen*)
participle subtype	FIRST PARTICIPLE (present), SECOND PARTICIPLE (past)
non-finite case	NOMINATIVE, GENITIVE, PARTITIVE, TRANSLATIVE, INESSIVE
non-finite number	SINGULAR, PLURAL
non-finite possessive suffix	THIRD PERSON SINGULAR
polarity	NEGATION
voice	ACTIVE, PASSIVE
mood	INDICATIVE, CONDITIONAL, IMPERATIVE, (POTENTIAL)
simplex tense	PRESENT, PAST
finite person-number	FIRST PERSON SINGULAR, SECOND PERSON SINGULAR, THIRD PERSON SINGULAR, FIRST PERSON PLURAL, SECOND PERSON PLURAL, THIRD PERSON PLURAL
Verb-chain general morphological features	

polarity	(AFFIRMATION), NEGATION
voice	(ACTIVE), PASSIVE
mood	INDICATIVE, CONDITIONAL, (IMPERATIVE)
person (finite+non-finite)	FIRST, SECOND, THIRD
number (finite+non-finite)	(SINGULAR), PLURAL
surface-syntax	CLAUSE-EQUIVALENT form, COVERT subject (implictly manifested agent)
Syntactic argument types	
AGENT, PATIENT, SOURCE, GOAL, MANNER, QUANTITY, LOCATION, TIME (as moment or period), DURATION, FREQUENCY, REASON+PURPOSE, CONDITION, META-ARGUMENT (clause-adverbial), NEGATIVE AUXILIARY, ADJACENT AUXILIARY, NON-ADJACENT NON-NEGATION AUXILIARY, (verb-chain-internal nominal) COMPLEMENT, (CO-ORDINATED CONJUNCTION), CO-ORDINATED VERB	
Semantic and structural subtypes of syntactic arguments and verb-chains	
AGENT	INDIVIDUAL, GROUP
PATIENT	HUMAN ← {INDIVIDUAL, GROUP}, ABSTRACTION ← {NOTION, STATE, ATTRIBUTE, TIME}, ACTIVITY, EVENT, INFINITIVE, PARTICIPLE, INDIRECT QUESTION, DIRECT QUOTE *että* ('that' subordinate clause)
MANNER	GENERIC, FRAME, POSITIVE (external evaluation), NEGATIVE, JOINT (activity), AGREEMENT
QUANTITY	MUCH, LITTLE
LOCATION	LOCATION (physical), GROUP, EVENT
TIME (as moment or period)	DEFINITE, INDEFINITE
DURATION	LONG, SHORT, OPEN, OTHER (fixed temporal reference)
FREQUENCY	OFTEN, AGAIN, OTHER ("non-often")
CO-ORDINATED VERB	MENTAL, ACTION
VERB-CHAIN (general semantic characteristics)	POSSIBILITY ← {POSSIBILITY (POSITIVE), IMPOSSIBILITY}, NECESSITY ← {NECESSITY (OBLIGATION), NONNECESSITY, FUTILITY}, EXTERNAL (cause), VOLITION, TEMPORAL, ACCIDENTAL
Extra-linguistic features	
QUOTATION (within newspaper text) MEDIUM: INTERNET newsgroup discussion (vs. NEWSPAPER text)	

As is evident in Table 1, there were in all quite a large number of contextual variables evident with substantial frequency in the research corpus. Of these, only a subset could be included in the multivariate analysis due to recommendations concerning the ratio of explanatory variables and outcome classes (synonyms) to the number of instances in the data (cf. Harrell 2001: 60-71). Consequently, semantic subtypes were included for only the most frequent syntactic argument types, and many feature variables were also lumped together, when possible and appropriate. The entire variable selection process, building upon univariate and bivariate statistical analyses, is presented in detail in Arppe (2008). In the end, 46 linguistic contextual feature variables were chosen for the "proper" full model (VI in Table 2), 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 classifications, and 6 semantic characterizations of the verb chains. This full model will be the "gold standard" (Harrell 2001: 98-99), against which we can then compare simpler models, incorporating different levels of linguistic analysis and their combinations, with varying degrees of overall complexity (i.e. Models I-V, VII-XI in Table 2). Furthermore, I am intrigued by what results might be produced with the entire variable set containing all the semantic and structural subtypes of the syntactic arguments identified in the corpus and satisfying a minimum frequency requirement ($n \geq 24$). Therefore, because the only real cost is computational, I will also try out such

an extended model, even at the risk of not setting the best example in the methodological sense. This extended model (VIII), when supplemented with extra-linguistic features (Model IX), largely conforms in its size and composition to the one used in Arppe (2007).

Table 2. Composition of the various features sets to be incorporated in the multivariate analysis models as explanatory variables.

Model index	**Feature set composition**	**Overall number of features**
I	Only node-specific morphological features	26
II	Verb-chain general morphological features (10) Node-specific features not subsumed by the verb-chain general features (17)	27
III	Syntactic argument types, *without* semantic and structural subclassifications	18
IV	Verb-chain general morphological features (10) Non-subsumed node-specific morphological features (17) Syntactic argument types (17) without their subtypes	44
V	Verb-chain general features (10) Most common semantic classifications of AGENTs and PATIENTs with their less frequent subtypes collapsed together (12) Other syntactic argument types *without* their subtypes (15)	37
VI	*"Proper" full model:* Verb-chain general morphological features (10) Semantic characterizations of verb-chains (6) Syntactic argument types alone (10) Syntactic argument types with selected or collapsed subtypes (20)	46
VII	Verb-chain general morphological features (10) Semantic characterizations of verb-chains (6) Syntactic argument types alone (10) Syntactic argument types with their subtypes (20) Extra-linguistic features (2)	48
VIII	*Extended full model*: Verb-chain general morphological features (10) Semantic characterizations of verb-chains (9) Syntactic argument types alone (5) All subtypes of syntactic arguments exceeding minimum frequency (38)	62
IX	Verb-chain general morphological features (10) Semantic characterizations of verb-chains (9) Syntactic argument types (5) All subtypes of syntactic arguments exceeding minimum frequency (38) Extra-linguistic features (2)	64
X	Extralinguistic features alone (2)	2
XI	Semantic characterizations of verb chains (6) Syntactic argument types alone (10) Selected or collapsed subtypes of syntactic arguments (20) (*excluding* any node-specific or verb-chain general morphological features)	36

Among various multivariate statistical methods, *polytomous logistic regression* analysis (see e.g. Hosmer and Lemeshow 2000: 260-287) appeared to be the most attractive approach. As a *direct probability model* (Harrell 2001: 217) polytomous as well as binary logistic regression yields probability estimates, corresponding to the expected proportions of occurrences, conditional on the values of the explanatory variables that have been selected for inclusion in the model. This characteristic fits well together with prior linguistic research (e.g., Featherston 2005, Bresnan et al. 2007, Arppe and Järvikivi 2007), from which we know that in practice individual

features or sets of features are *not* observed in corpora to be categorically matched with the occurrence (in a corpus) of only one lexeme in some particular synonymous set, or only one constructional variant, and no others. Rather, while one lexeme in a synonymous set, or one constructional alternative among the possible variants, may be by far the most frequent for some particular context, others do also occur, albeit with often a considerably lower relative frequency. Furthermore, with respect to the weighting of individual variables in polytomous logistic regression, the parameters associated with each variable have a natural interpretation in that they reflect the increased (or decreased) *odds* of a particular outcome (i.e. lexeme) occurring, when the particular feature is present in the context (instead of being absent), with all the other explanatory variables being equal. The exact meaning of the odds varies depending on which practical heuristic has been selected, and can concern e.g. a contrast with all the rest or with some baseline category.

There are a number of heuristics for implementing polytomous logistic regression, which are all based on the splitting of the polytomous setting into a set of dichotomous cases, to each of which a corresponding binary logistic regression model can then be applied and fitted either simultaneously or separately. These heuristics are presented and their characteristics discussed from the linguistic perspective in Arppe (2008, see also Frank and Kramer 2004). In order to get both lexeme-specific parameters for the selected contextual features, without having to select one lexeme as a baseline category, and probability estimates for the occurrences of each lexeme, I have found the *one-vs-rest* heuristic (Rifkin and Klautau 2004) as the most appealing of the lot. This methodological choice is facilitated by the observation that its performance does not significantly differ from that of the other heuristics, at least in the case of the studied phenomenon (Arppe 2007, 2008). The necessary statistical calculations were undertaken in the public-domain *R* statistical programming environment (R Core Development Team, 2007), using both ready-made functions (specifically `glm` for binary logistic regression incorporated in *R*'s `base` package) and functions written by myself. The latter were required for implementing the one-vs-rest heuristic, as well as statistic measures for the assessment of model performance, based on Menard (1995).

3. Results

Feature-wise odds for each of the THINK lexemes are already covered at length in Arppe (2007, 2008), so I will not discuss them any further here. I will rather shift the focus to the performance of different types of Models (I-XI), with varying levels of linguistic explanatory features and analytical complexity. Both the fit of the models and their prediction efficiency were evaluated using the entire research corpus as data (n=3404), the same which had been used to train the models, so the performance results should tentatively be considered somewhat optimistic. Nevertheless, validating the full 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), overall *Recall*=63.8% and 95% CI=(63.1%, 64.5%), $\lambda_{prediction}$=0.355 with 95% CI=(0.343, 0.368), and $\tau_{classification}$=0.479 with 95% CI=(0.468, 0.489) (see Arppe, 2008).

Of these four measures, R_L^2 is an indicator of how well a logistic regression model fits with the actual occurrences in the original data (Hosmer and Lemeshow 2000: 165-166). This is calculated as a comparison of the probabilities predicted by the model for each actually occurring outcome and the associated feature cluster, against the baseline probability for each outcome class, the latter which are simply the lexemes' overall proportions in the entire data. In comparison to the R^2 measure used in ordinary linear regression, R_L^2 does *not* tell us the proportion of variation in the data that a logistic regression model succeeds in explaining, but R_L^2 does allow us to compare the overall fit of different models with varying sets of explanatory variables on the same data. The three other measures concern efficiency in prediction (Menard 1995: 28-30). Firstly, *Recall* tells us how often overall a prediction is correct, based in the case of the one-vs-rest heuristic on a prediction rule of selecting per each context the lexeme receiving the highest probability estimate. The *Recall* measures presented here are an aggregate of the lexeme-wise *Recall* values, which in the case of the above full model are quite divergent, favoring *ajatella* with a mean *Recall* of 85.40%, in comparison to the respective values of 45.1% for *miettiä*, 49.5% for *pohtia*, and 46.0% for *harkita*. Secondly, $\lambda_{prediction}$ is a measure for the *proportionate reduction of prediction error*, which tells us how much better the model performs over the baseline strategy of always picking the most frequent outcome class (i.e. the mode, in this case *ajatella*). Thirdly, $\tau_{classification}$ is the measure for *proportionate reduction of classification error*, which on top of instance-wise prediction accuracy considers how well the model is able to replicate the overall distribution of outcome classes in the original data, in this case the relative lexeme frequencies.

As can be seen in Table 3, increasing the number of feature categories and levels in linguistic analysis quite naturally has a positive impact on how much of the occurrences of the selected THINK lexemes can be accounted for. Starting at the simplest end, node-specific morphology (Model I), and somewhat surprisingly even if supplemented with verb-chain general morphological features (Model II), as well as extra-linguistic features alone (Model X), appear to have roughly equal (and low) explanatory power both in terms of fit with the original data as well as their added value in prediction. The *Recall* levels for these three models (I: 47.15%, II: 47.71% and X: 47.21%) do not substantially rise above the proportion of the most frequent THINK lexemes, *ajatella*, in the research corpus, being 1492/3404=43.8%. This is in fact reflected in the measures concerning the reduction of prediction error with $\lambda_{prediction}$ ranging 0.059-0.060-0.059, which indicate a minimal improvement in the results over always predicting the most frequent outcome class. In contrast, the measures for the reduction of classification error with these models are already clearly higher, with $\tau_{classification}$ ranging at 0.239-0.240-0.247, but among all the models considered here these values rank nevertheless as the lowest.

Table 3. The descriptive and predictive properties of the various types of Models (I-XI) with different compositions of explanatory variables, based on the single-fit training and testing of each model with the one-vs-rest heuristic data on the entire data (*n*=3404).

Model index	**Recall (%)**	R_L^2	$\lambda_{prediction}$	$\tau_{classification}$
I	47.15	0.094	0.059	0.239
II	47.71	0.100	0.069	0.247
III	50.18	0.098	0.113	0.282
IV	56.82	0.180	0.231	0.378
V	63.04	0.288	0.342	0.468

VI	64.60	0.313	0.370	0.490
VII	65.57	0.325	0.387	0.504
VIII	65.60	0.325	0.388	0.504
IX	65.80	0.337	0.391	0.507
X	47.21	0.057	0.060	0.240
XI	63.10	0.292	0.343	0.468

Syntactic argument types alone (Model III), without any of their semantic and structural subtypes and excluding all morphological features, fare already slightly better. The fit with the original data is roughly equal to that achieved with the node-specific and verb-chain general morphological features (Models I-II), and almost twice the corresponding value for extralinguistic features (Model X). As *Recall* with Model III increases to above the half-way-mark, the measures of prediction and classification error improve also accordingly, with $\lambda_{prediction}$ almost doubling in value in contrast to Models I-II and X; for $\tau_{classification}$ the absolute improvement is of a similar magnitude but lesser in relative terms. When morphological features concerning the entire verb-chain and the node are combined with syntactic argument types (Model IV), the performance overall notches up noticeably. Now, the fit with the original data at R_L^2=0.180 is twice that of the morphological or syntactic arguments types alone (Models I-III), and over three times the level reached with extralinguistic features (Model X). Whereas *Recall* increases moderately to only 56.82%, especially the reduction of prediction error in comparison to syntactic argument types alone (Model III) roughly doubles, and also classification error reduces considerably, with $\lambda_{prediction}$=0.231 and $\tau_{classification}$=0.378.

If we further supplement the morphological and syntactic argument features with the semantic and structural classifications of the two most common and important arguments in the case of the THINK lexemes, namely their AGENTs and PATIENTs (Model V), the results in terms of the descriptive fit of the model with the original data or prediction accuracy all improve again visibly. While *Recall* increases to 63.04%, the other measures grow less modestly by roughly one-third, as now R_L^2=0.288, $\lambda_{prediction}$=0.342 and $\tau_{classification}$=0.468. In contrast, adding further the subtypes for MANNER and TIME (as a moment or period) arguments as well as the semantic classifications of verb-chains incorporated in the full Model (VI) does not continue the improvement of the performance of the models at the same rate. Now, though descriptive fit has yet grown somewhat to R_L^2=0.313, on the predictive side *Recall* has increased by only one percent to 64.6%, while the reduction of prediction error is modestly up at $\lambda_{prediction}$=0.370 and $\tau_{classification}$=0.490.

The most complex model with the extended semantic classifications (Model VIII, with as many as 16 more semantic subtypes of syntactic arguments in comparison to Model VI) produces but quite minute improvements, with R_L^2=0.325, *Recall*=65.6%, $\lambda_{prediction}$=0.388 and $\tau_{classification}$=0.504. Thus, it would appear that we are approaching some sort of upper limit, seemingly around a level of two-thirds accuracy in prediction, as to what can be achieved with the types of quite conventional linguistic analysis features applied in this study, concerning morphology, syntax and semantics within the immediate sentential context. A similar conclusion was earlier noted in Arppe (2007) with a slightly differently selected extended variable set. Furthermore, dropping out the proper morphological verb-chain general features altogether but retaining the semantic characterizations of verb-chains and combining these with the

syntactic arguments as well as those among their semantic subtypes selected for the full Model (VI), amounting to the feature set in Model XI, results in a surprisingly small drop in performance, as R_L^2=0.292 with a *Recall*=63.1%, $\lambda_{prediction}$=0.343 and $\tau_{classification}$=0.468. Thus, the linguistic information coded in the morphological features, whether on the node-verb or the associated verb-chain in general, would appear to an essential extent be already incorporated in the syntactic and semantic argument structure. This is supported by the fact that the mean odds for morphological features, when incorporated into a model together with syntactic arguments and their semantic subtypes as well as overall semantic characterizations of verb-chains, are considerably smaller in comparison to those for these other feature categories (Arppe, 2008).

As these results are clearly less than the performance levels achieved by Gries (2003b, *Recall*=88.9%, canonical *R*=0.821) and Bresnan et al. (2007, *Recall*=92%), even if achieved in simpler dichotomous settings, one possible avenue for improvement would be to add entirely new linguistic analysis categories such as longer-distance discourse factors, as was done in these prior studies. However, the inclusion of the two extralinguistic features selected in this study, indicating the medium of usage (newspaper vs. Internet newsgroup discussion, and quoted fragments vs. body text), yield only small improvements of around one percent-unit in magnitude for the various performance measures. This is apparent for both Model VII, for which the performance measures are R_L^2=0.325, *Recall*=65.57%, $\lambda_{prediction}$=0.387 and $\tau_{classification}$=0.504, as well as for Model IX, in which case the corresponding values are R_L^2=0.337, *Recall*=65.8%, $\lambda_{prediction}$=0.391 and $\tau_{classification}$=0.507. These results correspond in absolute terms to 33 more correctly classified lexeme selections with Model VII in comparison to Model VI, but only 7 with Model IX in comparison to Model VIII.

Furthermore, similar, less than perfect levels of prediction accuracy (54%[1]) have been reached for the even more complex 6-way prediction of synonymous Russian TRY verbs, using the simultaneously fit multinomial heuristic with a baseline category. In this particular case, the explanatory variables have consisted of the semantic properties of their subjects and the following infinitives as well as Tense-Aspect-Mood (TAM) marking on the TRY verbs themselves (personal communications from Dagmar Divjak 4.12.2007, 16.5.2008 and 19.5.2008). This would suggest that the performance levels reached in this study would not at all be exceptionally poor or low. By contrast, Inkpen and Hirst (2006: 25-27, see also Inkpen 2004: 111-112) achieved over 90 percent accuracy in correctly selecting a synonym from several multiple-lexeme sets. This would indicate that the choices can in fact be highly precisely modeled, but this requires explanatory variables indicating 1) "nuances" such as denotational microdistinctions, 2) the speaker's intention to express some attitude, and 3) the sought-after style. These are not necessarily explicitly evident in the immediate sentential context nor easily amenable to accurate automated extraction (Edmonds and Hirst 2002: 128, cf. Hanks 1996: 90, 97).

[1] In the validation of this model, the jack-knife estimate was 50.8%. Furthermore, splitting 100 times randomly the entire data sample of 1351 instances into training sets of 1000 instances and testing sets with the remaining 351 instances yielded a mean correct classification rate of 49%, with a standard deviation of 2.45% (Personal communication from Dagmar Divjak 16.5.2008).

4. Discussion

The current performance plateau may result from technical restrictions related to the application of the one-vs-rest heuristic in particular, and on the basis of the similarities in the performance of all the heuristics demonstrated in Arppe (2008), of polytomous logistic regression in general, to the more complex, multiple-outcome setting in this study. This may also result to some extent from the exclusion of interaction terms among the explanatory variables included in all the Models I-XI presented above, due to restrictions set by the size of the available data and its outcome frequencies. But this might also reflect genuine synonymy, or at least some extent of interchangeability in at least some contexts, which the current analysis variables cannot (possibly: can never) get an exact hold of (cf. Gries 2003b: 13-16). Even more radically we may interpret such (varying degrees of) interchangeability as evidence rather for inherent variability in language, following Bresnan (2007).

The underlying premises of logistic regression analysis, i.e. assuming relative proportions of occurrence rather than categorical selections, suggest that we should not focus only at the maximum probabilities assigned for each instance (which according to the classification rule for the one-vs-rest heuristic, i.e. $arg_{Lexeme}\{max[P(Lexeme|Context)]\}$, determine the lexeme predicted per each instance). Rather, we should expand our scrutiny to the entire spectrum of probabilities estimated for each outcome (i.e. *Lexeme* ~ *L*) in a particular context (~ *C*). Indeed, as we can see in Figure 1, the maximum probability assigned (using Model VIII) for any lexeme in any context rarely approaches the theoretical maximum $P(L|C)=1.0$, and the predictions are practically categorical in only 258 (7.6%) instances for which $P_{max}(L|C)>0.90$. On the contrary, the mean maximum probability per all instances and contexts is only $\bar{x}(P_{max}[L|C])=0.636$, while the overall span of maximal values is as broad as (0.28, 1.00), and even the 95% *Confidence Interval* is very wide at *CI*=(0.369, 0.966). The lower-ranked instance-wise probability estimates have similar overall characteristics of intermediate-level means and broad ranges. The second-highest probability estimates per instances have a mean $\bar{x}(P_{max\text{-}1}[L|C])=0.244$, with an overall range of (0.000, 0.490) and a 95% *CI*=(0.026, 0.415), and the third-highest (i.e. second-lowest) probability estimates have a mean $\bar{x}(P_{max\text{-}2}[L|C])=0.096$, with an overall range of (0.000, 0.307) and a 95% *CI*=(0.000, 0.241). Even the minimum probability estimates clearly keep some distance from zero as their mean $\bar{x}(P_{min}[L|C])=0.043$, even though their overall range is (0.000, 0.212) as well as 95% *CI*=(0.000, 0.144). Nevertheless, as many as 764 (22.4%) of the minimum estimated probabilities per instance are practically nil with $P_{min}(L|C)<0.01$. However, turning this the other way around, for 2640 (77.6%) instances in the entire data the minimum estimated probability $P_{min}(L|C)\geq 0.01$. This latter case represents an expected possibility of occurrence at least once every hundred times or even more often in a similar context for *all four* THINK lexemes.

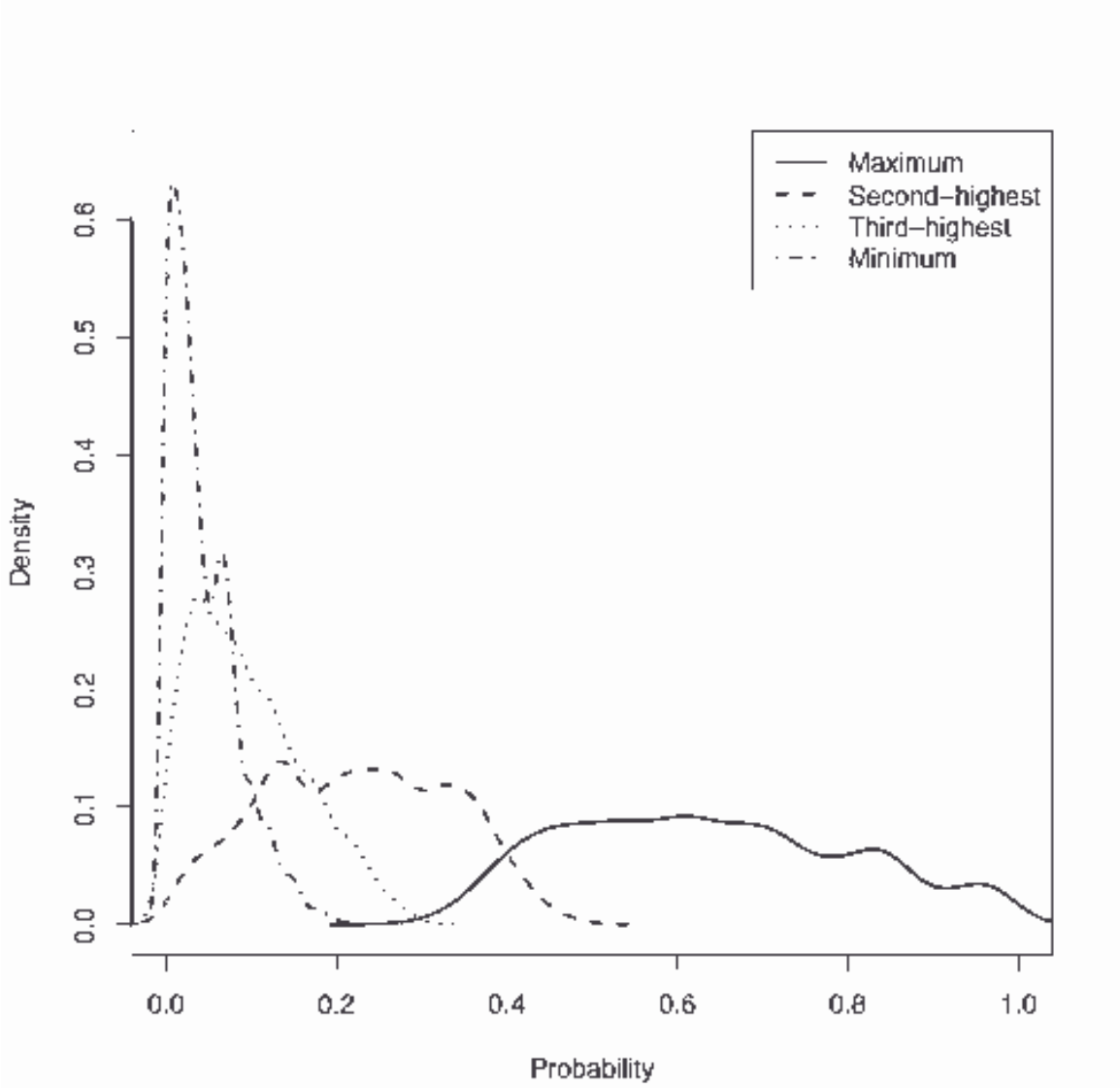

Figure 1. Densities of the distributions of the estimated probabilities by rank order for all instances in the data (*n*=3404).

Looking at the instance-wise estimated probabilities as a whole, in 64 (1.9%) instances all four estimates are $P(L|C)\geq 0.15$, indicating relatively equal values for all lexemes, and in 331 (9.7%) instances all four are $P(L|C)\geq 0.10$. Discarding always the minimum value, in 303 (8.9%) cases the remaining three higher-ranked probability estimates are all $P(L|C)\geq 0.2$, and in as many as 1436 (42.2%) cases $P(L|C)\geq 0.10$. Narrowing our focus only to the two topmost-ranked lexemes per instance, in 961 (26.2%) cases both probability estimates are $P(L|C)\geq 0.3$, and for as many as 150 (4.4%) cases both $P(L|C)\geq 0.4$. The contextual settings associated with these last-mentioned instances would be prime candidates for fully or partially synonymous usage within the selected set of THINK lexemes, as their joint probabilities would indicate high mutual interchangeability. In sum, these distributions of instance-wise probability estimates for all four THINK lexemes suggest that, to the extent these probabilities even approximately represent the proportions of actual occurrences in given contexts, very few combinations of contextual features are associated with categorical, exception-less outcomes. On the contrary, quite a few of the contexts can realistically have two or even more outcomes, though preferential differences among

the lexemes remain to varying extents (cf. Hanks 1996: 79). Lastly, let us assume for the sake of argument that a theory about language consists simply of two parts: 1) the fundamental components of which language is considered to consist and with which language can be comprehensively analyzed, and 2) the rules or regularities concerning how these components interact and are allowed to combine into sequences. If we accept that the contextual features used in this study are good and satisfactory representatives of a theory of language, these results certainly support Bresnan's (2007) probabilistic view of the relationship between language usage and the underlying linguistic system.

Zooming in on individual sentences in the research corpus (Table 3), we can observe various scenarios of how the entire estimated probability space (with $\sum P[L|C]=1.0$) can be distributed among the THINK lexemes on the basis of the selected features manifested in each context. Firstly, the probability distribution may approach categorical, exception-less choice, so that only one of the lexemes is assigned in practice the maximum possible probability $P(L|C)\approx 1.0$, while the rest receive none (exemplified by sentence #1 in Table 3). However, such a scenario applies to as few as 7.6% of the sentences in the research corpus. Secondly, selectional situations for some contexts may inherently incorporate variation so that one lexeme is clearly preferred in such circumstances, receiving the highest probability, but one or more of the others may also have a real though more occasional chance of occurring to a varying degree (e.g. sentence #2 in Table 3). This can also be observed to result (as a logical consequence of the premises of logistic regression modeling) in individual instances of actual usage for which the originally selected lexeme is not the one which has been assigned the highest probability estimate (e.g. sentence #3 in Table 3). Lastly, we can observe cases in which all four lexemes are estimated to have approximately equal probability with respect to the observable context (e.g. sentence #4 in Table 3). Such instances with close-to-equal estimated probabilities of occurrences could be considered as candidate examples of "genuine" synonymy and full interchangeability in context for the entire selected set of four THINK lexemes. These quite sensible scenarios, identified on the basis of manual inspection of the original data, can in fact be verified by applying statistical techniques such as hierarchical agglomerative clustering (HAC) to systematically arrange and group the entire set of lexeme-wise probability distribution estimates available for all instances (n=3404) in the data (Arppe 2008).

Table 3. A small selection of sentences from the research corpus with varying distributions of estimated probabilities for the four THINK lexemes; maximum probability in boldface (e.g. **0.5**); probability assigned to actually occurring lexeme under-lined (e.g. 0.5); pertinent feature variables as subscripts next to the appropriate word (or head in the case of a phrase/clause)

#/(Features)	**Sentence**
#1(7) P(*ajatella*|*Context*$_1$)=**1** P(*miettiä*|*Context*$_1$)=0 P(*pohtia*|*Context*$_1$)=0 P(*harkita*|*Context*$_1$)=0	*Miten*$_{MANNER+GENERIC}$ ***ajattelit***$_{INDICATIVE+SECOND,\ COVERT,\ AGENT+INDIVIDUAL}$ *erota*$_{PATIENT+INFINITIVE}$ *mitenkään jostain SAKn umpimielisistä luokka-ajattelun kannattajasta?* [3066/politiikka_9967] 'How did you **think** to differ at all from some dense supporter of class-thinking in SAK?'
#2 (7) P(*ajatella*|*Context*$_2$)=0.018	*Vilkaise*$_{CO-ORDINATED\_VERB(+MENTAL)}$ *joskus*$_{FREQUENCY(+SOMETIMES)}$ *valtuuston esityslistaa ja*

P(*miettiä*|*Context*$_2$)=**<u>0.878</u>** P(*pohtia*|*Context*$_2$)=0084 P(*harkita*|*Context*$_2$)=0.020	***mieti***$_{(IMPERATIVE+)SECOND,\ COVERT,\ AGENT+INDIVIDUAL}$ *monestako*$_{PATIENT+INDIRECT\_QUESTION}$ *asiasta sinulla on jotain tietoa.* [2815/politiikka_728] 'Glance sometimes at the agenda for the council and **think** on how many issues you have some information.'
#3 (8) P(*ajatella*|*Context*$_3$)=0.025 P(*miettiä*|*Context*$_3$)=0.125 P(*pohtia*|*Context*$_3$)=<u>0.125</u> P(*harkita*|*Context*$_3$)=**0.725**	*Tarkastusviraston mielestä*$_{META}$ *tätä ehdotusta*$_{PATIENT+ACTIVITY}$ *olisi*$_{CONDITIONAL+THIRD,\ COVERT}$ *syytä*$_{VERB\_CHAIN+NECESSITY}$ ***pohtia*** *tarkemmin*$_{MANNER+POSITIVE}$. [766/hs95_7542] 'In the opinion of the Revision Office there is reason to **ponder** this proposal more thoroughly.'
#4 (8) P(*ajatella*|*Context*$_4$)=**<u>0.301</u>** P(*miettiä*|*Context*$_4$)=0.272 P(*pohtia*|*Context*$_4$)=0.215 P(*harkita*|*Context*$_4$)=0.212	*Aluksi harvemmin, mutta myöhemmin tyttö alkoi viettää öitä T:n luona ja vuoden tapailun päätteeksi* $P_{AGENT+INDIVIDUAL}$ *sanoi, että voisi*$_{CONDITIONAL+THIRD,VERB\text{-}CHAIN+POSSIBILITY,COVERT}$ ***ajatella*** *asiaa*$_{PATIENT+ABSTRACTION(<NOTION)}$ *vakavamminkin*$_{MANNER+POSITIVE}$. *(SFNET)* [50/ihmissuhteet_8319] '... P said that [she] could **think** about the matter more seriously [perhaps]'

Scrutinizing the actual linguistic contexts in the example sentences in Table 3 with at least some degree of dispersion among the lexeme-wise probability estimates (i.e. #2, #3, and #4, as well as similar cases in the entire research corpus, see Arppe, 2008), I find it difficult to identify any additional contextual features or essentially new feature categories, which would allow us to distinguish among the lexemes or select one over the rest. Here, I restrict my consideration to features which pertain to current, conventional models of morphology, syntax and semantics, and which concern the immediate sentential context. Rather, it seems that the semantic differences between using any of the THINK lexemes in these example sentences are embedded and manifested in the lexemes themselves. Moreover, these distinctions would appear to be of the kind that do not and would not necessarily have or require an explicit manifestation in the surrounding context and argument structure. That is, the selection of any one of the THINK lexemes in these sentences each emphasizes some possible, though slightly distinct aspect or manner of THINKing. Nonetheless, all such aspects could be mostly fully conceivable and acceptable as far as concerns the constraints set by the surrounding linguistic structure. In this, the relevant discriminatory selective characteristics would concern features outside the traditional linguistic domain, i.e. expressed attitude, emotion and style. These correspond to the "nuances" which Inkpen and Hirst (2006: 1-4) have found accurate in reduplicating which of the various near-synonymous alternative lexemes (with the tested sets comprising more than two synonyms) have actually been used (Inkpen and Hirst 2006: 26-27). Such shades of meaning, which could be considered to incorporate the implications and presuppositions discussed by Hanks (1996), cannot in the most cases be resolved on the basis of the immediate sentence context alone. However, they might be deduced from prior passages in the same text from which the particular sentence is taken, or from previous related texts in the same discussion thread, or on the basis of extralinguistic knowledge about the context or even concerning the participant persons in the linguistic exchange (cf. Hanks 1996: 90, 97).

5. Conclusions

In conclusion, the observed general upper limit to *Recall* in prediction, at approximately two-thirds, or 64.6-65.6% to be exact, of the instances in the research corpus, as well as an in-depth scrutiny of the sentences with lexemewise dispersion among the estimates of probability, can be viewed to represent the explanatory limits of linguistic analysis which can be reached within the immediate sentential context and applying the conventional descriptive and analytical apparatus based on currently available linguistic theories and models (cf. Gries 2003b: 13-16). Looking from the other angle of the estimated probabilities for lexical outcomes, given a set of contextual features, the results indicate that there exists for the most part substantial and tangible variation with respect to which lexemes can actually occur in the close-to-same contexts. In fact, for 77.6% of the sentences in the research corpus the estimated expected probabilities are for all four lexemes at least $P(Lexeme|Context)>0.01$. The closer inspection of not only sentences with roughly equal estimates of probability for all four lexemes but also those with non-categorical preferences for one or two of the lexemes would suggest that such variation in context is both common and acceptable. Furthermore, it seems that any distinctive features there may be are not explicitly evident in the immediate sentential context, but rather pertain to stylistic attitudes and intended shades of expression that the speaker/writer writer wishes to convey (belonging to the intermediate stylistic/subconceptual level in the clustered model of lexical choice by Edmonds and Hirst 2002). More generally, these results support a probabilistic notion of the relationship between linguistic usage and the underlying linguistic system, akin to that presented by Bresnan (2007). Few choices are categorical, given the known context (feature cluster) that can be analytically grasped and identified. Rather, most contexts exhibit various degrees of variation as to their outcomes, resulting in proportionate choices on the long run. Nevertheless, these results should be corroborated with other types of linguistic evidence, for instance experimentation, such as e.g. Bresnan (2007) and Gries (2003b) have done.

Acknowledgements

I am grateful for insightful comments and feedback provided to me by Martti Vainio, Lauri Carlson, Dagmar Divjak and Simo Vihjanen with respect to the interpretation of the results, though the burden of accurate representation of all matters in this paper rests solely on myself.

Corpora

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 – Center for Scientific Computing, 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 – Center for Scientific Computing, Finland. Available on-line at URL: http://www.csc.fi/kielipankki/

References

Arppe, Antti
2002 The usage patterns and selectional preferences of synonyms in a morphologically rich language. In: Morin, Annie and Pascale Sébillot (eds.), *JADT-2002. 6th International Conference on Textual Data Statistical Analysis*, 13-15.3.2002, Vol. 1, 21–32. Rennes: INRIA.

2006 Complex phenomena deserve complex explanations. *Quantitative Investigations in Theoretical Linguistics* (QITL2) Conference, Osnabrück, Germany, 1-2.6.2006, 8–11. Available on-line at URL: http://www.cogsci.uni-osnabrueck.de/~qitl/

2007 Multivariate methods in corpus-based lexicography. A study of synonymy in Finnish. In: Davies, Matthew; Paul Rayson; Susan Hunston; and Pernilla Danielsson (Editors), *Proceedings from the Corpus Linguistics Conference (CL2007)*, July 28-30, 2007, Birmingham, UK. Available on-line at: URL: http://www.corpus.bham.ac.uk/corplingproceedings07/

2008 Univariate, bivariate and multivariate statistical methods in corpus-based lexicography – A study of synonymy. [Ph. D. dissertation]. Publications of the Department of General Linguistics, University of Helsinki, No. 44. Available on-line at: URN: http://urn.fi/URN:ISBN:978-952-10-5175-3.

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

Atkins, Beryl T. S. and Beth Levin
1995 Building on a Corpus: A linguistic and lexicographical look at some near-synonyms. *International Journal of Lexicography* 8 (2): 85–114.

Bresnan, Joan
2006 Is syntactic knowledge probabilistic? Experiments with the English dative alternation. *Pre-proceedings of the International Conference on Linguistic Evidence*, Tübingen, 2-4 February 2006, 3–10. Tübingen: Sonderforschungsbereich 441, University of Tübingen.

2007 Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In: Featherston, Sam and Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*. Series: Studies in Generative Grammar. Berlin: Mouton de Gruyter.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina and R. Harald Baayen

2007 Predicting the Dative Alternation. In: Boume, G., Kraemer, I. and J. Zwarts (eds.), *Cognitive Foundations of Interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

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

Church, Kenneth, William Gale, Patrick Hanks and Douglas Hindle
1991 Using Statistics in Lexical Analysis. In: Zernik, Uri (ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, 115-–164. Hillsdale: 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 (visited 29.5.2007) and URL: http://www.connexor.com/demo/doc/enfdg3-tags.html (visited 5.6.2007).

Divjak, Dagmar
2006 Ways on Intending. Delineating and Structuring Near-Synonyms. In: Gries , Stefan Th. and Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics*. Vol. 2: The syntax-lexis interface, 19–56. Berlin: Mouton De Gruyter.

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

forthcoming Clusters in the mind? Converging evidence from Near-synonymy in Russian.

Edmonds, Philip and Graeme Hirst
2002 Near-synonymy and Lexical Choice. *Computational Linguistics* 28 (2): 105–144.

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

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

Frank, Eibe and Stefan Kramer
2004 Ensembles of Nested Dichotomies for Multi-Class Problemes. *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada.

Gries, Stefan Th.

2003a *Multifactorial analysis in corpus linguistics: a study of particle placement*. London: Continuum.

2003b Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics*, Vol. 1, 1–27

Hanks, Patrick
1996 Contextual Dependency and Lexical Sets. *International Journal of Corpus Linguistics*, 1 (1): 75–98.

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

Hosmer, David W., Jr., and Stanley Lemeshow
2000 *Applied Regression Analysis* (2nd edition). New York: Wiley.

Inkpen, Diana
2004 Building a Lexical Knowledge-Base of Near-Synonym Differences. Ph. D. dissertation, Department of Computer Science, University of Toronto.

Inkpen, Diana and Graeme Hirst
2006 Building and Using a Lexical Knowledge-Base of Near-Synonym Differences. *Computational Linguistics* 32 (2): 223–262.

Jantunen, Jarmo H.
2001 Tärkeä seikka ja keskeinen kysymys. Mitä korpuslingvistinen analyysi paljastaa lähisynonyymeistä? [Important point and central question. What can corpus-linguistic analysis reveal about near-synonyms] *Virittäjä* 105 (2): 170–192.

2004 *Synonymia ja käännössuomi: korpusnäkökulma samamerkityksisyyden kontekstuaalisuuteen ja käännöskielen leksikaalisiin erityispiirteisiin* [Synonymy in translated Finnish. A corpus-based view of contextuality of synonymous expressions and lexical features specific to translated languages]. Ph. D. dissertation. (University of Joensuu Publications in the Humanities 35). Joensuu: University of Joensuu.

Järvinen, Timo and Pasi Tapanainen
1997 *A Dependency Parser for English*. TR-1, Technical Reports of the Department of General Linguistics, University of Helsinki.

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

Menard, Scott
1995 Applied Logistic Regression Analysis. (Sage University Paper Series on Quantitative Applications in the Social Sciences 07-106.) Thousand Oaks: Sage Publications.

Miller, George A.

1990 Nouns in WordNet: a lexical inheritance system. (revised August 1993). *International Journal of Lexicography*, 3 (4): 245–264. Available on-line at: URL: ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps

Pajunen, Anneli
2001 *Argumenttirakenne: Asiaintilojen luokitus ja verbien käyttäytyminen suomen kielessä* [Argument structure: the classification of states-of-affairs and the behavior of verbs in Finnish]. (Suomi 187.) Helsinki: Suomalaisen Kirjallisuuden Seura.

R Development Core Team.
2007 *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL: http://www.R-project.org.

Rifkin, Ryan and Aldebaro Krakatau
2004 In Defense of One-Vs-All Classification. *Journal of Machine Learning Research*, 5 (January): 101-141.

Tapanainen, Pasi and Timo Järvinen
1997 A non-projective dependency parser. In: *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, D.C., April 1997, Association of Computational Linguistics, 64–71.

Yang, Charles
2008 The great number crunch. *Journal of Linguistics*, 44, 205–228.