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

The purpose of this paper is to present a case study elucidating how multivariate models can be interpreted to shed light on the nature of the use and choice among lexical and structural alternatives in language, more specifically near-synonyms, and the underlying explanatory factors. By multivariate models I imply both the use of multiple linguistic variables from a range of categories, instead of only one or two, in order to study and explain some linguistic phenomenon, as well as the use of multivariate statistical methods such as polytomous logistic regression.

In the modeling of lexical choice among semantically similar words, specifically near-synonyms, it has been shown in (mainly) lexicographically motivated corpus-based studies of actual lexical usage that semantically similar words differ significantly on the syntactic-semantic level as to the 1) lexical context, the 2) syntactic structures which they form part of, and the 3) semantic classification of some particular argument(s). However, Gries (2003a) has demonstrated that univariate explanations are not at all sufficient; rather, lexical or syntactic choices made by speakers/writers are determined, and can thus be explained considerably more satisfactorily, by a plurality of factors, representing different linguistic categories, in interaction. Furthermore, Bresnan (to appear) has suggested that the outcomes of such combinations of variables, i.e. selections in context, are generally speaking probabilistic, even though the individual choices in isolation are discrete. That is, the workings of a linguistic system, represented by the range of variables according to some theory, and its resultant usage are not in practice categorical, following from exception-less rules, but exhibit degrees of potential variation which becomes evident over longer stretches of linguistic usage. Moreover, both Gries (2003b) and Bresnan (to appear, et al. 2007) have shown that there is evidence for such probabilistic character both in natural language use in corpora as well as language judgments in experiments, and that these two sources of evidence are convergent. However, these studies have focused on dichotomous syntactic alternatives, namely verb-particle placement and the dative alternation in English. 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'. The data consists of corpus material from the foremost Finnish newspaper as well as Finnish Internet newsgroup discussion fora, containing altogether 3404 occurrences of the studied lexemes, in roughly equal proportions from both two sources. Each observed instance as well as its context was analyzed by hand morphologically, syntactically and semantically (see Arppe, to appear). Whereas Gries (2003b) used Linear Discriminant Analysis, I will employ in the statistical analyses logistic regression, similar to Bresnan et al. (2007), since its results as a direct probability model are more attractive in that they have natural interpretations (cf. Arppe, to appear). All variables are binary, indicating whether a particular linguistic or extra-linguistic feature or feature cluster occurs or applies in a given context or not.

Table 1. Performance of the polytomous logistic regression models for the THINK lexemes with various sets of explanatory variables [number of variables per each category in brackets]

Variable sets/ Performance	**Recall (%)**	R_L^2	$\lambda_{prediction}$
Node-specific morphology [27]	47.8	0.095	0.071
Node-specific [16] + verb-chain morphology [11]	48.1	0.101	0.075
Node-specific [16] + verb-chain morphology [8] + syntactic argument types [17]	55.0	0.160	0.198
All syntactic argument types (alone) [18]	49.8	0.084	0.107
Verb-chain morphology [11] + syntactic argument types [6] + semantic subtypes [43]	65.3	0.328	0.381
Verb-chain morphology [11] + syntactic argument types [6] + semantic subtypes [43] + extra-linguistic features [3]	65.8	0.340	0.391
Extra-linguistic features alone [3]	46.9	0.059	0.055

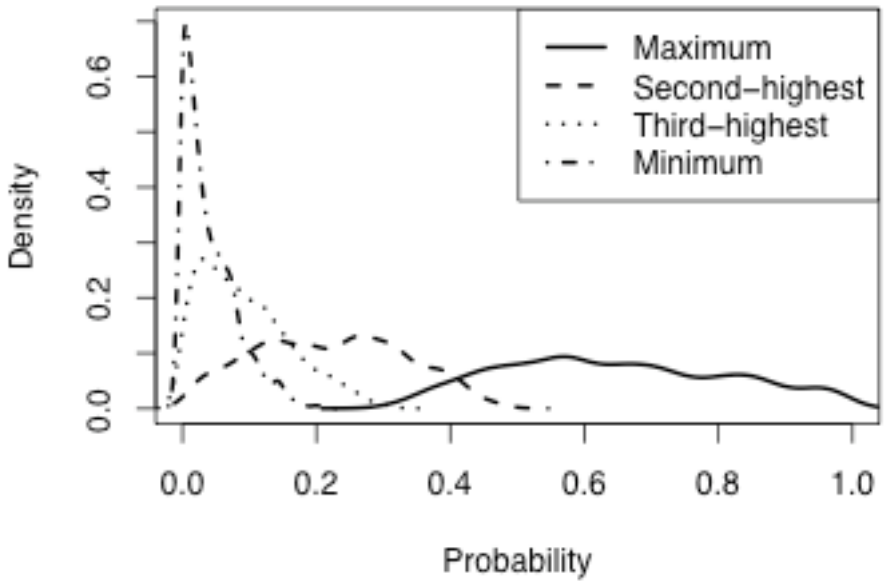
Table 2. Average importance of the different categories of explanatory variables on the basis of means of the statistically significant odds-ratios in the full polytomous regression model incorporating all variable types (mean values including also non-significant cases in parentheses).

Variable subtype	**Mean odds (including non-significant cases)**
Node-specific and verb-chain morphology	2.6002 (1.6450)
Syntactic argument types (alone)	3.0231 (1.7387)
Syntactic argument types + semantic subtypes	4.2571 (2.1064)
Extra-linguistic features	1.8061 (1.5579)

As can be seen in Table 1, increasing the number of feature categories in linguistic analysis quite naturally has a positive impact on how much of the studied phenomenon can be accounted for. Interestingly, either node-specific morphology or syntactic argument structure alone have roughly equal (and low) explanatory power. Combining these two levels of analysis together notches the results up noticeably, as

does supplementing semantic classifications. However, as was noted in Arppe (to appear), adding more intricacy on the semantic level does not appear to significantly improve the results. Therefore, it would seem that the approximately two-thirds (65.8%) accuracy in predicting the actual choice in the corpus might be the upper limit that can be reached on the basis of conventional morphological, syntactic and semantic analysis restricted to the immediate sentential context. As these results are clearly less than the performance levels achieved by Gries (2003b, *Recall*=88.9%, canonical R=.821) and Bresnan et al. (2007, *Recall*=92%), even if achieved in simpler dichotomous settings, one possible solution would be to include longer-distance discourse factors as was done in these prior studies; however, the addition of a few extra-linguistic variables indicating medium and repetition had no substantial effect here (amounting in practice only to an addition of 19 correctly classified selections). Yet, using extra-linguistic features alone reached almost the same prediction accuracy level as morphological features by themselves, suggesting that extra-linguistic characteristics are already to some extent incorporated in the linguistic features proper. Nevertheless, we can in the final full model evaluate what are the average weights of the various variable categories. As can be seen in Table 2, syntactic arguments coupled with a semantic classification are clearly the most distinctive group of features, followed at a distance by syntactic argument types alone, and then morphological features pertaining to both the node-verb and the possibly associated verb-chain, while the extra-linguistic features have the least impact.

Figure 1. Probabilities of lexemes per each context.



The current performance plateau may result from the more complex, multiple-outcome setting in this study, but it might also reflect genuine synonymy, or at least some extent of interchangeability in at least some contexts, which the current analysis variables cannot (possibly never) get an exact hold of (cf. Gries 2003b: 13-16). In fact, when we look at the distributions of the maximum probabilities assigned for all the contexts in the studied data (Figure 1), we can note that less than 10% of all contexts receive a higher probability than P(lexeme|Context)=0.90, with the average maximum probability being only 0.65 (s.d. 0.17). Moreover, the second highest probabilities in the same contexts are substantial, with a mean value of 0.22 (s.d. 0.11), while a large part of minimum probabilities are also clearly more than nil, averaging 0.040 (s.d. 0.039). These results would seem supportive of Bresnan's probabilistic view of the relationship between linguistic usage and the underlying linguistic system, as well as suggestive of the explanatory limits of linguistic analysis, though they should be corroborated with experimentation (as Bresnan and Gries have done).

References

Arppe, A. (to appear). Multivariate methods in corpus-based lexicography. A study of synonymy in Finnish. Fourth Biennial Corpus Linguistics 2007 Conference, 28-30.7.2007, Birmingham, United Kingdom.

Bresnan, J. (to appear). Is syntactic knowledge probabilistics? Experiments with the English dative alternations. In: S. Featherston and W. Sternefeld eds. Roots: Linguistics in search of its evidential base. Series: Studies in Generative Grammar [SSG] 96. Mouton de Gruyter, Berlin, Germany.

Bresnan, J., A. Cueni, T. Nikitina, and H. Baayen. (2007). Predicting the Dative Alternation. In G. Boume, I. Kraemer, and J. Zwarts eds. Cognitive Foundations of Interpretation, pp. 69-94.

Gries, S. Th. (2003a). Multifactorial analysis in corpus linguistics: a study of Particle Placement. London, Continuum Press, New York, New York.

Gries, S. Th. (2003b). Towards a corpus-based identification of prototypical instances of constructions. Annual Review of Cognitive Linguistics, 1:1-27.