Morphology in the extreme: Labial Reduplication in Hungarian

Labial reduplication (LR) in Hungarian is a semi-productive morphological template with a diminutive/hypocoristic function, which consists in creating a copy of the base with a labial consonant in initial position; some examples are *cica*-[m]*ica* 'cat.DIM', *csiga*-[b]*iga* 'snail.DIM', *Julcsi*-[b]*ulcsi*, hypocorism for *Julia*. While there are detailed accounts of similar phenomena in other languages (see e.g. Southern 2005 for Yiddish), the Hungarian pattern has not been systematically described before. The lack of any useful descriptive material is partly due to the fact that LR exhibits a high degree of variability and is difficult to tease apart from a more general tendency to form rhyming compounds. In this talk, we analyse data collected from a 1.5 billion word corpus of Hungarian (Halácsy et al., 2004) and attempt to isolate and describe this pattern with the help of machine learning algorithms and statistic methods. We also discuss the implications of the data for theories of morphology and phonology.

The corpus search that serves as the basis of our analysis presented two major difficulties. First, while it seems intuitively clear that LR is a distinct pattern in Hungarian word-formation, there are many similar forms where the reduplicant has a non-labial initial consonant (e.g. *angyal*-[k]*angyal* 'angel.DIM'). Second, there is a considerable number of forms that seem to be reduplicated, but are in fact the result of compounding (e.g. *csillog*-[v]*illog* 'shines and flashes'). To avoid any potential biases resulting from our preconceptions about the pattern, a very general search algorithm was used, which extracted all items of the form $O_1\{\dots\}_i$-$O_2\{\dots\}_i$ (O = onset and $O_1 \neq O_2$) from the corpus. The resulting items were grouped as shown in Table 1. Groups 1a and 1b contain forms created through reduplication of either the first or the second element (e.g. 1a *cica*-[m]*ica* from *cica*, 1b *ici*-[p]*ici* 'tiny' from *pici* 'small'), group 2 forms created through compounding, group 3 iconic forms where both elements are meaningless (e.g. *locs*-[p]*ocs* 'splish splash') and group 4 loanwords (e.g. *super-duper*). Although these forms are created through a number of different processes, it is reasonable to assume that LR could affect the selection of forms within each group as a product-oriented schema (cf. Bybee 2001).

The pattern of LR can be expected to have a strong effect within groups 1a, 1b and potentially 3, and a weaker effect within groups 2 (where the choice of the two elements is strongly influenced by semantic considerations) and 4 (where there is interference from other languages). This hypothesis receives confirmation from two sources. First, the relative frequency of formations where the second element is labial-initial is very high in groups 1a, 1b and 3, and relatively low in groups 2 and 4, as shown in Figure 1. Second, the labial pattern is subject to a certain degree of phonetic conditioning (see below), that is, the choice of the initial consonant is not entirely random. Consequently, a classification algorithm that exploits generalisations within the data should be able to make better predictions about the behaviour of items within groups where the labial pattern is stronger. We used the Tilburg Memory-Based Learner (Daelemans et al., 2007) to test this prediction; the classification accuracy was considerably higher within groups 1a, 1b and 3 than it was within groups 2 and 4. This is shown in Figure 2. It is important to note that group 2 is not entirely free from the effects of the labial pattern: there is a high proportion of rhyming compounds with labials, which suggests that speakers are more likely to create compounds that fit the pattern of LR. As for the phonological conditioning, two potential factors were tested: the quality of the first and the second onset of the base. The influence of these two factors was evaluated by performing a multidimensional scaling of the data (Cox & Cox, 2001) based on the proportions of different behaviours associated with different qualities in these positions (see Figures 3 and 4). The first onset shows an anti-repetition effect: [m]-initial words take a [b]-initial reduplicant, and [p]/[b]-initial words an [m]-initial reduplicant; moreover, there is a strong tendency for words with an [l] in their onset to take [f]. The second onset shows more systematic conditioning: voiceless consonants attract a higher proportion of [m]-initial reduplicants than voiced ones.

We believe that these results argue for models of morphological/phonological competence that can incorporate the effects of individual exemplars and token-frequency (cf. Bybee 2001). The fine-grained conditioning described above is hard to account for in categorical models, while it can be explained quite naturally in exemplar-based models. Moreover, information about token-frequency significantly increases the performance of categorisation algorithms. The interaction of LR and compounding can also only be accounted for in models which allow for some overlap between different patterns.

LABEL	STEM 1	STEM 2	MECHANISM
1a	+	−	reduplication
1b	−	+	reduplication
2	+	+	compounding
3	−	−	iconic formation
4	−	−	borrowing

Table 1: Grouping of $O_1\{\ldots\}_i$-$O_2\{\ldots\}_i$ forms; columns 2 and 3 indicate whether the first/second element of the formation appears as an independent stem in Hungarian.

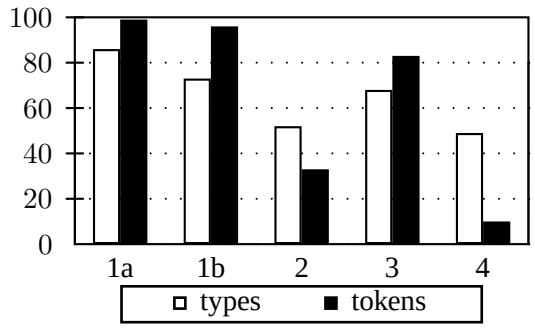

Figure 1: Relative frequency of labials

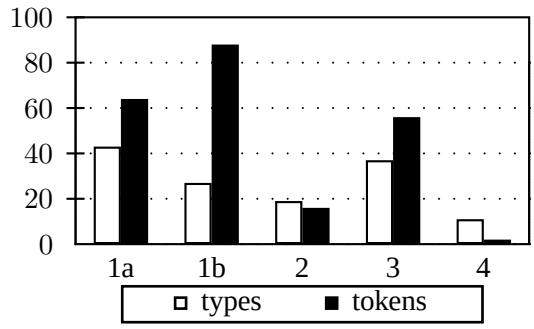

Figure 2: TiMBL classification accuracy

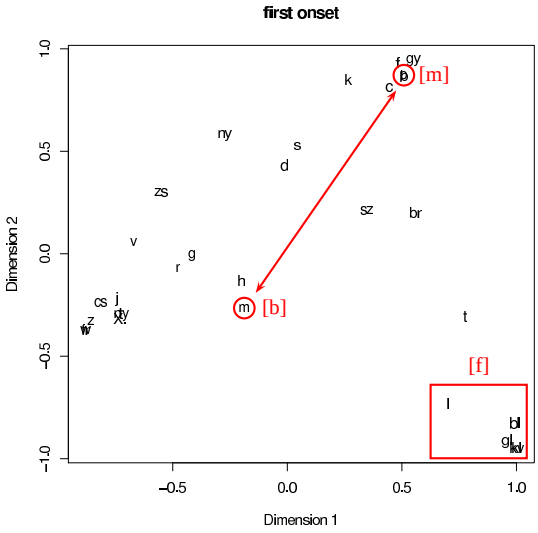

Figure 3: Multidimensional scaling for 1st onset

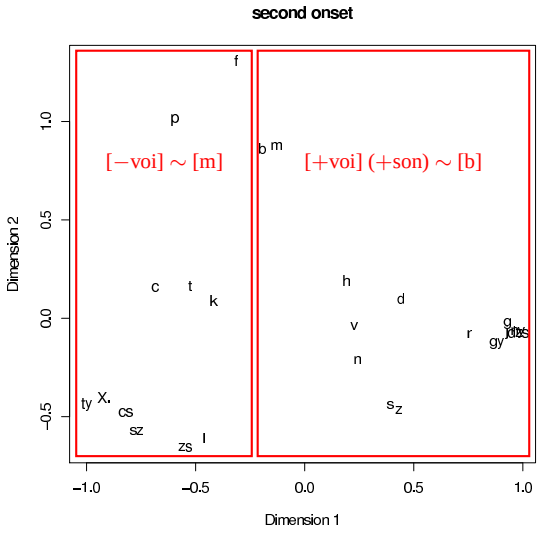

Figure 4: Multidimensional scaling for 2nd onset

References

Bybee, J. L. (2001). *Phonology and language use*. Cambridge University Press, Cambridge.

Cox, T. F. & M. A. A. Cox (2001). *Multidimensional scaling*. Chapman & Hall, Boca Raton, FL.

Daelemans, W., J. Zavrel, K. van der Sloot & A. van den Bosch (2007). TiMBL: Tilburg Memory Based Learner, version 6.1, Reference Guide. ILK Research Group Technical Report Series no. 07-07.

Halácsy, P., A. Kornai, L. Németh, A. Rung, I. Szakadát & V. Trón (2004). Creating open language resources for Hungarian. In: Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004).

Southern, M. R. V. (2005). *Contagious couplings: Transmission of expressives in Yiddish echo phrases*. Praeger, Westport, CT & London.