7 Morphology in the extreme: echo-pairs in Hungarian*

Márton Sóskuthy
University of Edinburgh

1 Introduction

Echo-pair formation is a cross-linguistically well-attested process, which consists in the creation of a word-like unit composed of two nearly identical parts, only differing in their initial consonants and/or their vowels (see Southern 2005 for a similar use of the term). Some examples are *itsy-bitsy*,[1] *fancy-shmancy*, *splish-splash* (English), *Schorle-morle* 'spritzer' (German), *et-met* 'meat or something' (Turkish; Southern 2005: 60), *paampu-kiimpu* 'snakes and other such creatures' (Tamil; Keane 2001: 1). As can be seen from the examples, echo-pair formation encompasses a broad range of morphological, phonological and semantic patterns. Thus, such formations can be created through reduplication of a base form (e.g. *fancy-shmancy*), compounding (e.g. *rumble-jumble*), borrowing (e.g. *füle-müle* 'nightingale' in Hungarian < *Philomela* in Greek) and spontaneous invention (e.g. *plick-plock*; Thun 1963: 50). As for the phonological makeup of echo-pairs, the component forms sometimes differ only in their initial consonants, sometimes only in their vowels and sometimes in both (forms where the two components are identical do not count as instances of echo-pairs under the present definition). The semantics of these constructions also varies widely: the most common functions associated with echo-pairs include diminutive, hypocoristic, and dismissive. Moreover, the wide range of variation in the function and form of echo-pairs is often attested not just cross-linguistically, but even within a single language.

Somewhat surprisingly, echo-pairs — in Hungarian and in general — have received very little attention in the phonological and morphological literature. In particular, there is a complete lack of works discussing the possible theoretical consequences of the phenomenon, and there are hardly any treatments that offer a systematic analysis of the data that they present. For example, Thun (1963) provides

*This paper has grown out of a presentation with Bálint Feyér, Péter Rácz and Dániel Szeredi. I am greatly indebted to them for their help in compiling the data set and for their comments on the analysis. I would also like to thank Ádám Nádasdy for his insightful review of the paper.

[1]Throughout this paper, the two components of echo-pairs will be separated by a hyphen both in spelling and in phonetic transcriptions, regardless of the spelling conventions of the source language.

an extensive set of examples from English taken from a range of different periods and dialects, but does not make any broader generalisations regarding the form of the observed variants. Similarly, Apor (1906), Simonyi (1907) and Szikszainé Nagy (1993) present an impressive array of cross-dialectal data from Hungarian, but do not identify and analyse the various phonological trends observable in their data sets.

I believe that the reason for the shortcomings of previous treatments of echo-pairs is to be found in the nature of the data. As it has been noted above, echo-pairs tend to exhibit an unusually wide range of variation, which is likely due to the fact that the morphological pathways leading to such formations tend to be only partially conventionalised. As a result, the data sets from various languages often seem rather chaotic, with few clearly identifiable tendencies. This makes them unsuitable for analysis in traditional descriptive and theoretical frameworks based on categorical phenomena. Moreover, the descriptive works referred to above collapse data from several different dialects into a single data set. Since most theoretical approaches focus on the competence of a single speaker (cf. generative approaches such as Chomsky and Halle 1968 and Prince and Smolensky 1993) or the language use of a given community (cf. sociophonetic approaches such as Labov 1994), it is not clear what conclusions could be drawn from such a mixed set of variants. Therefore, it is clear that echo-pairs call for a different approach both in terms of data collection and analysis.

This paper focusses on a relatively small and well-defined subset of the problems described above, namely the phonological and morphological aspects of echo-pairs in Hungarian. Importantly, forms involving vowel changes are not dealt with, and the semantics and cross-linguistic aspects of the pattern will also not be discussed in any detail. As it will be seen, Hungarian echo-pairs exhibit the full range of variation described above. However, it is possible to isolate a smaller group of words in Hungarian which behave more systematically both in phonological and morphological terms: echo-pairs with a labial-initial second component. Here are a number of representative examples: *cica-mica* ‘cat.DIM’, *csiga-biga* ‘snail.DIM’, *Ancsi-Pancsi* ‘Anna.HYPO’.[2]

The paper endeavours to answer the following three questions related to echo-pairs:

[2]Hungarian examples are presented in standard orthography. The following letter-to-sound correspondences should be noted: <c> [ts], <cs> [tʃ], <gy> [ɟ], <ny> [ɲ], <s> [ʃ], <sz> [s], <ty> [c], <zs> [ʒ]. The rest of the consonants have their standard IPA values; IPA transcriptions will be provided when the quality of the vowels is relevant.

1. What types of methods can be used to collect data about echo-pairs in Hungarian? (methodology)
2. What phonological tendencies can be identified in echo-pairs? (description)
3. What theoretical implications does echo-pair formation have? Is it possible to go beyond the level of description? (theory)

This three-part distinction is also reflected in the general layout of the paper. Thus, in Section 2, I take up a number of methodological issues: I show how a large set of echo-pairs can be extracted from a corpus of Hungarian and how these forms can be arranged to allow for more in-depth analysis. Then, in Section 3, I use a variety of statistical and computational methods to identify phonological trends in the resulting data set, providing a more general description of the phenomenon through an analysis of the distribution of initial consonants in the second component of echo-pairs. Finally, in Section 4, I look at the theoretical implications of echo-pair formation, and propose that the observed trends can best be described as the result of the productive use of morphological schemata in the sense of Bybee (2001). Section 5 concludes the paper with a summary of its main points.

2 Data collection

The analysis presented in this paper relies on a set of forms extracted from the 600 million word Hungarian Webcorpus (Halácsy et al. 2004). The main reasons for choosing a corpus search over other methods of data collection are as follows. Controlled elicitation tasks are necessarily restricted to a set of forms preselected by the experimenter; this makes them unsuitable for the purposes of the present paper, which intends to explore the full range of variation in echo-pair formation in Hungarian. Traditional informant-based methods would also result in a skewed data set, as they encourage the production of forms that the informants consider interesting or peculiar, and which do not necessarily represent their competence. A corpus search, on the other hand, is likely to yield examples of actual language use, and a reasonably large corpus can also be expected to contain a representative sample of the echo-pairs that occur in Hungarian.

The Hungarian Webcorpus is particularly well-suited to the study of echo-pairs, since all the material in the corpus comes from the Internet and is therefore often written in a relatively informal register. Echo-pairs are usually restricted to playful and intimate contexts due to the semantics of the template, and are therefore more likely to occur in a corpus containing informal text samples than in a corpus

consisting mainly of relatively formal ones. The size of the corpus also ensures that it exemplifies the full range of variation in echo-pair formation.

It should be noted that the Hungarian Webcorpus contains text samples from speakers of a variety of different dialects, which means that the corpus-search method does not solve the problem of collapsing different language varieties into a single data set. However, since the source of the material is the Internet, it can be assumed that the speakers address themselves to a regionally non-specific audience. It has been observed that the speech style of a given speaker is affected by the audience (see e.g. Bell 1984), which means that it is unlikely that dialect-specific forms dominate in the corpus. Moreover, the corpus-search method also provides information about the token frequency of individual items. Since expressions specific to a given dialect are likely to have a lower frequency, they will have a relatively weaker influence on the results of the analysis (provided that it is based on token-frequency, or that infrequent forms are excluded).

Since the Hungarian Webcorpus does not currently have a phonetically transcribed version, the corpus search had to be based on written representations, which might be seen as a problem given that this paper focusses on the phonological aspects of echo-pairs. However, Hungarian spelling is relatively predictable in most cases, and the majority of the graphemes are in a one-to-one correspondence with their phonemic values, which means that written forms can be used where phonemic representations would normally be required. Indeed, the decision to rely on written forms has not caused any difficulties either during data collection or analysis.

The template used for the corpus search can be described as follows:

(1) $O_1\{\dots\}_i O_2\{\dots\}_i$, $O_1 \neq O_2$
O: a string of consonant characters (an onset; O_1 might be null)
$\{\dots\}_i$: a string of at least three characters starting with a vowel

That is, a corpus search using this the template returns all strings consisting of two identical parts where only the initial onsets differ. It was necessary to impose a lower limit on the length of the individual parts, as a corpus search without this restriction would return too many irrelevant forms (including all disyllabic words where the syllable rhymes are the same, such as *Miki* 'Nick' < *Miklós*, *lámpám* 'my lamp' < *lámpa* + *-(V)m*). To illustrate the scope of the template, here are a few examples for matching and non-matching forms:

(2) match: *cica-mica*, *Isti-Pisti*, *blicc-hicc*
no match: *ciróka-maróka*, *izeg-mozog*, *nyam-nyam*

Note that the template requires full identity of the characters following the onset, which means that forms showing vowel changes are excluded from the data set (see some of the examples in (2)).

The raw data set produced by the corpus search initially contained more than 4,000 word forms. However, more than half of these forms turned out not to be useful for the purposes of the present paper: some of these were foreign words (e.g. *backpack*), some suffixed forms (e.g. *ásatása*) and some snippets of code and unintelligible sequences that have not been removed from the original corpus (e.g. *mdashdmdash*). This left an overall 2,048 word forms, which were collapsed into 1,446 types (keeping an overall frequency count for each type).

These forms can be divided into four major groups on the basis of the word formation processes they exemplify: reduplicated forms (e.g. *cica-mica* 'cat.DIM'), rhyming compounds (e.g. *csillog-villog* 'is very clean' from *csillog* 'glistens' and *villog* 'flashes'), iconic formations where no base form can be identified (e.g. *csiri-biri* 'hocus pocus', where neither *csiri* nor *biri* exist as independent words)[3] and loanwords (e.g. *blackjack* [blɛgdʒɛkk]). A simple criterion can be applied to tease these different morphological formations apart from each other: forms where only one of the component parts occur independently in the language are likely to be the result of reduplication, whereas forms where both or none of the component parts occur independently are cases of compounding and iconic formation/borrowing, respectively (see Thun 1963: 10 for a similar criterion). Table 1 illustrates this grouping, and introduces a further distinction within the group of reduplicated forms based on the order of the base and the reduplicant.

PART 1	PART 2	MECHANISM	EXAMPLE	GLOSS
+	−	reduplication	***cica***-*mica*	'cat.DIM'
−	+	reduplication	*ici*-***pici***	'very small'
+	+	compounding	***csillog*-*villog***	'is very clean'
−	−	iconic formation	*dumm-bumm*	'rumbling sound'
−	−	borrowing	*black-jack*	'blackjack'

Table 1: *Types of word-formation processes based on the criterion of independence. Columns 1 and 2 indicate whether the first/second element of the formation appears as an independent stem in Hungarian. The forms typeset in bold occur independently in the language.*

Table 2 shows a few sample entries from the resulting database.

[3]The term 'iconic' is used since the relationship between the meaning and the sound shape of these forms is usually not entirely arbitrary. See Jakobson (1965) for a similar use of the term.

SPELLING	MORPHOLOGY	SYLLABLES	INIT	FREQ
cica-mica	reduplication	[ts,i,=] [ts,ɒ,=]	[m]	366
nyuszi-gyuszi	compound	[ɲ,u,=] [s,i,=]	[ɟ]	2
ecc-pecc	iconic form	[=,ɛ,tsː] [=,=,=]	[p]	28

Table 2: *Sample entries from the echo-pair database*

Note that the forms are stored in a syllabified form, where each syllable consists of an onset, a nucleus and a coda (thus, [ts,i,=] corresponds to a syllable with [ts] as its onset, [i] as its nucleus and no coda). Since the present study focusses on the distribution of initial consonants in the second component of echo-pairs, the onset of the second component is represented separately (this is indicated in the fourth column of Table 2). The database also contains the token frequency of each form.

3 Data analysis

This section presents a phonological analysis of the data set described in the previous section. The main emphasis is on the predictability of the initial consonant of the second component, which is henceforth referred to simply as the *behaviour* of the echo-pair (e.g. *cica-mica* exhibits behaviour [m] and *csiga-biga* behaviour [b]). The behaviour of echo-pairs is studied as a function of two main classes of variables: the phonological makeup of the rest of the word (cf. the third column in Table 2) and the morphological type of the echo-pair (cf. the second column in Table 2). It will be shown that these two classes of variables are strongly interrelated with each other, in the sense that the behaviour of certain morphological types shows clearer phonological conditioning. This observation serves as the basis of the structure of the section. Thus, I first investigate the behaviour of echo-pairs created through reduplication and find that a number of relatively clear phonological patterns can be identified within this group (3.1). In the second part of the section, I show that while some of these phonological tendencies can also be observed in echo-pairs created through other word formation processes, they have a much weaker effect outside the group of genuine reduplicated forms (3.2).

3.1 The behaviour of reduplicated forms

Echo-pairs created through reduplication (e.g. *cica-mica* and *ici-pici*) have a number of special properties which make them particularly well-suited to the study of the

phonological conditioning of echo-pair formation. This is mainly due to the fact that this is the only case where echo-pair formation comes close to more pedestrian morphological processes: reduplication takes a base-form and modifies it in a given way to create a form that fits the general pattern of echo-pairs, similarly to more familiar cases of affixation and templatic morphology. The reduplicant does not exist independently in the language (e.g. the string *mica* in *cica-mica* is not a Hungarian word), and the resulting echo-pair does not have any iconic properties, which means that it is unlikely that this morphological process is affected by any factors other than the phonological makeup of the base. This is clearly not the case for either compounding, where the choice of the two components is likely to be affected by semantic considerations as well, or iconic formations, where the iconicity of the echo-pair might impose additional constraints on its behaviour.

As it has been noted in Section 2, there are, in fact, two different patterns of reduplication that can result in echo-pairs. One of these consists in the addition of phonological material at the right edge of the base (as in the case of *cica-mica*, where *mica* is added to *cica*) and the other in the addition of phonological material at the left edge of the base (as in the case of *ici-pici*, where *ici* is added to *pici*). Although these two word formation processes could, in principle, differ in terms of phonological conditioning, the present analysis collapses them into a single group. While this might be seen as problematic, the results presented below demonstrate that the class of reduplicated forms exhibits a high degree of phonological conditioning as a whole, despite the fact that the directionality of reduplication is ignored.

Before turning to the effect of the phonological composition of the base on the behaviour of reduplicated echo-pairs, it will be useful to take a closer look at the distribution of behaviours within this morphological class. As it has been noted in Section 1, echo-pairs with a labial-initial second component have a special status in Hungarian. This is particularly clear within the reduplicated class, where a remarkably high proportion of the forms shows a labial behaviour: more than 98% of all the reduplicated tokens belongs to this behavioural class.[4] This is to be expected in light of the remarks above: if reduplicated forms reflect the phonological tendencies in echo-pair formation without any distortions resulting from non-phonological factors, and forms with a labial behaviour enjoy a privileged position among echo-pairs, we should not be surprised to find a high number of such forms within the reduplicated group. Figure 1 shows a more detailed summary of the distribution of behaviours

[4]The following discussion focusses on token frequency to the exclusion of type frequency. The reason for this restriction is as follows. The data set contains a relatively high number of *hapax legomena* and other infrequent forms, many of which exemplify slightly unusual production patterns (e.g. *ejnye-nejnye* 'tut-tut' which occurs only once in the data set, as opposed to *ejnye-bejnye*, which occurs 4,087 times). The use of token frequencies results in infrequent forms receiving less weight in the analysis, which prevents them from significantly skewing the results.

among reduplicated forms (behaviours occurring less than 50 times in the corpus are omitted).

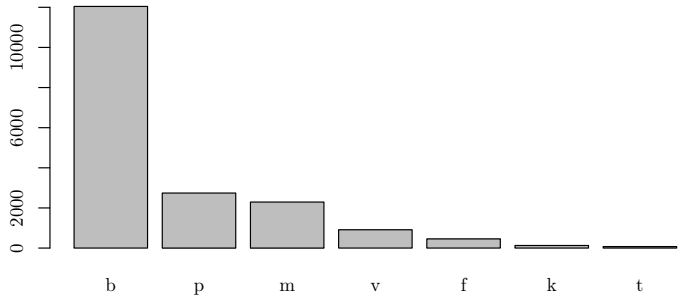

Figure 1: *The distribution of behaviours among reduplicated forms.*

The five most frequent behavioural patterns are all labial (which exhausts the range of labials in Hungarian) and [b], [p] and [m] together account for almost 91% of all the behaviours within this group. Therefore, the rest of this section focusses mainly on the latter three behaviours.

Further analysis of the patterns proceeded as follows. A preliminary inspection of the data set revealed a number of suggestive tendencies, which were used to formulate hypotheses about the phonological conditioning of the behaviour of reduplicated echo-pairs. These hypotheses were then tested through a combination of machine-learning and statistical methods to be described below. The hypotheses are listed below:

(3) THE FIRST ONSET:

a. [m]-initial first component → behaviour: [b]
b. [p]/[b]-initial first component → behaviour: [m]
c. vowel-initial first component → behaviour: [p]/[b]

(e.g. *mogyi-bogyi*, *puszkó-muszkó*, *Ancsi-Pancsi*)[5]

(4) THE SECOND ONSET:

a. voiced second onset → behaviour: [b]
b. voiceless second onset → behaviour: [m]/[p]

(e.g. *édi-bédi*, *cica-mica*)

The hypothesis testing procedure was based on the following basic principles. The data set can be divided into several subsets based on the values the features listed in the third column of Table 2 can take on. For example, it is possible to induce a partitioning of the data set based on the first onset of its first component

(henceforth referred to simply as 'the first onset'), which will result in several groups including [ts]-initial forms, [tʃ]-initial forms, [t]-initial forms, and so on (e.g. *cica-mica*, *csiga-biga*, *tündi-bündi* 'lovely'). Each of these groups will show a particular distribution of behaviours, and — if the hypotheses are correct — these distributions will be at least partly predictable on the basis of the feature values the partitioning is based on.

In principle, these distributions could be compared directly and the results used to confirm or reject the hypotheses above. As an illustration, consider Table 3, which only shows a small subset of the possible behaviours and the feature values.

x	$P([\text{m}]\|x)$	$P([\text{b}]\|x)$	$P([\text{p}]\|x)$
O_1=[k]	0.58	0.16	0.04
O_1=[t]	0.27	0.27	0.04
O_1=[tʃ]	0.29	0.47	0.07

Table 3: *The distribution of behaviours* [m]*,* [b] *and* [p] *among echo-pairs with* [k]*,* [t] *and* [tʃ] *as their first onset. $p(y|x)$ stands for the conditional probability of a given behaviour within a given group, which is calculated as its proportion within the group.*

The table shows that while [k]-initial forms attract a [m]-type behaviour, [tʃ]-initial forms favour [b]; [t]-initial forms seem to be intermediate in this respect (these tendencies are not related to the ones listed in ((3)) and ((4))). When looking at the full range of feature values and behaviours, this method requires the inspection of approximately 10-20 values and the comparison of the proportions of 10-20 possible behaviours for each of them. Since this procedure is exremely time-consuming and not sufficiently illuminating, a different method will be used for analysing the data set.

The groups can also be compared on a more abstract level by using a distance function to obtain a numeric measure of how different they are in terms of their behaviour. Ideally, such a measure would place [k] and [tʃ]-initial forms relatively far from each other and [t] somewhat closer to each of them. One possible way of calculating these distances is by using the Modified Value Difference Metric (MVDM) as described in Daelemans et al. (2007). The MVDM uses the following formula to calculate the distances between individual groups (where v_1 and v_2 are different feature values, B_i is a given behaviour and n is the total number of behaviours):

$$(5) \qquad d(v_1, v_2) = \sum_{i=1}^{n} |P(B_i|v_1) - P(B_i|v_2)|$$

That is, the MVDM simply computes how different the two groups are in terms of each possible behaviour and sums the results. In the present case, the MVDM was used to create a matrix of distance values for each pair of groups.

The most straightforward interpretation of these distance values is to imagine a space of behaviours in which the different groups occupy different locations. The hypotheses in (3) and (4) could be tested by comparing the locations of the relevant groups in this space (since forms that behave similarly will be close together and forms that behave differently will be far apart). Such a set of locations can be constructed and visualised with the help of a method called *multidimensional scaling* (Cox and Cox 2001), which maps a distance matrix to a set of low-dimensional coordinates (two-dimensional in the present case). It may be easier to understand the principles of multidimensional scaling through a somewhat less abstract example. Imagine a set of cities (e.g. Edinburgh, Budapest and Tromsø), where the exact location of the cities is not known, only the distance between each pair (1786 km for Edinburgh-Budapest, 1873 km for Tromsø-Edinburgh and 2467 km for Budapest-Tromsø). Multidimensional scaling can be used to create a set of coordinates which specify the actual location of the three cities on a map of Europe (after rescaling and rotating the coordinates). Exactly the same operation can be performed for a given partition of the set of reduplicated echo-pairs.

Figure 2 shows a two-dimensional visualisation of the distribution of behaviours among reduplicated echo-pairs as a function of their first onset (based on types; forms with a token frequency of less than 5 were excluded to avoid the distorting effects of low-frequency items).[6]

The figure also uses colour to visualise an additional type of information, namely the proportion of behaviours [m], [b] and [p] (which are represented by red, blue and green, respectively). Intermediate hues represent feature values which attract a mixed behavioural pattern. The degree to which these two types of representation (colour and location) are correlated is remarkable: the values of the first onset are arranged in a triangle whose three corners are each associated with one of the three colours; intermediate locations also correspond to intermediate hues. Since the calculation of the coordinates involves all the behaviours (as opposed to the colours, which represent only [m], [b] and [p]), this suggests that the result of the multidimensional scaling is mostly determined by the three most frequent behaviours.

A quick look at Figure 2 is enough to confirm all the three hypotheses about the influence of the first onset on the behaviour of reduplicated echo-pairs in (3).

[6]The characters t', d' and n' stand for [c], [ɟ] and [ɲ], respectively. The rest of the characters have their standard SAMPA values; the equality sign indicates a null onset.

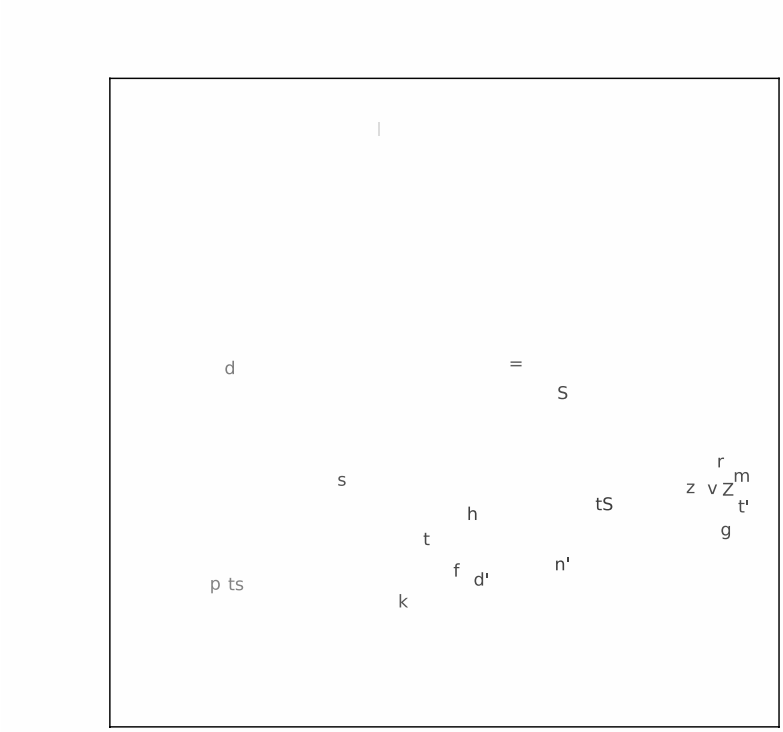

Figure 2: *The distribution of behaviours among reduplicated echo-pairs as a function of their first onset.*

First, [m] is in the lower right corner of the triangle, which is clearly associated with behaviour [b] (cf. (3a)). Second, [p] is in the lower left corner of the triangle, where forms with behaviour [m] reside (cf. (3b)).[7] Finally, vowel-initial forms (represented by the equality sign) are located between the two corners corresponding to behaviours [b] and [p] (cf. (3c)). The first two of these tendencies can also be interpreted as examplifying a more general pattern of dissimilation: labial-initial forms tend to have a second component that starts with a different labial consonant. It could be objected that the data set does not actually contain forms where the two components are identical due to the structure of the search template, and is thus not suitable for investigating patterns of dissimilation (as evidence against dissimilation is excluded by definition). However, it is quite telling that the patterns noted above seem to be conventionalised: [m]-initial forms are only found with behaviour [b] but

[7] Unfortunately, no [b]-initial forms were left after the removal of infrequent forms, which means that this part of the hypothesis cannot be tested.

not [p], and [p]-initial forms only with behaviour [m] but not [b]. The fact that each different type of labial-initial echo-pair in the data set selects a single behaviour — together with the informal observation that forms where the two components are identical do not seem to have the same semantic properties as forms where they differ — suggests that dissimilation is a valid interpretation of the observed tendencies.

The influence of the second onset on the behaviour of reduplicated echo-pairs is illustrated in Figure 3.

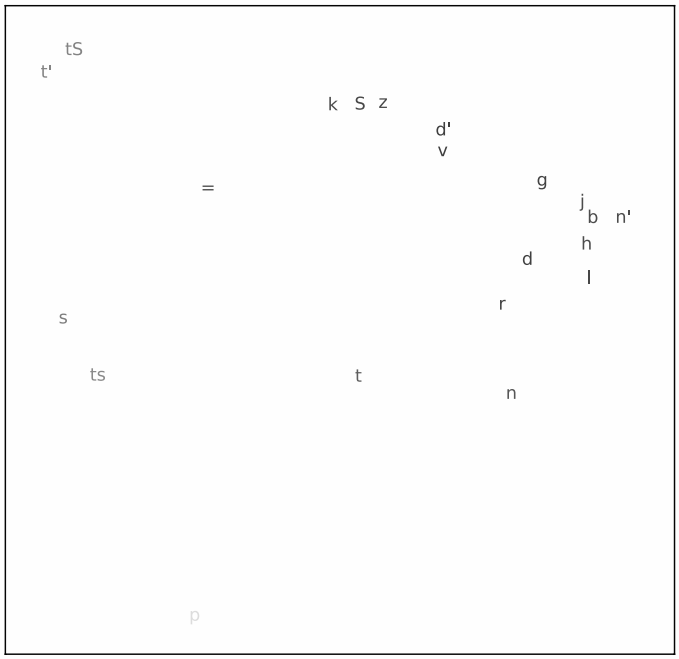

Figure 3: *The distribution of behaviours among reduplicated echo-pairs as a function of their second onset.*

Once again, the feature values are arranged in a triangle whose corners correspond to the three most frequent behavioural patterns. Interestingly, the position of the different consonantal values along the x-axis seems to be correlated with their voicing: voiceless consonants are found on the left hand side of the diagram and voiced consonants on the right hand side. This supports the hypotheses in (4): voiceless consonants in the second onset attract behaviours [p] and [m], as

opposed to voiced consonants, which attract behaviour [b]. It should be noted that this tendency is not as strong as those observed for the first onset: there are quite a few values which occupy an intermediate position and cannot be clearly associated with a single behavioural pattern. However, the values closer to the corners of the triangle all behave in accordance with the hypotheses in (4). The only exception is [h], which, however, has long been noted for its ambiguous behaviour with respect to voicing: it is the only voiceless obstruent that does not undergo voicing before a voiced obstruent (see e.g. Siptár and Törkenczy 2000). The relationship between the voicing of the second onset of the echo-pair and its behaviour can also be interpreted as a case of assimilation: forms with a voiced consonant exhibit a voiced behaviour, as opposed to forms with a voiceless consonant, which exhibit a voiceless or passively voiced behaviour.[8] It should be noted that such patterns of assimilation are extremely rare: the acoustic cues of voicing hardly extend beyond the consonant they belong to, and therefore do not typically trigger patterns of long-distance assimilation (Hansson 2004; Blevins and Garrett 2004).

3.2 Compounds, iconic forms and loanwords

The investigation of echo-pairs formed through reduplication has shown that echo-pair formation is subject to a relatively high degree of phonological conditioning when there is no interference from non-phonological factors. However, the methods used in the analysis of reduplicated forms cannot be straightforwardly extended to compounds, iconic formations and loanwords. The reasons for this are as follows. Semantic considerations play an important role in the choice of the two components of echo-pairs in compounds, and can be expected to override the phonological patterns observed in the previous section. Similarly, the sound symbolic aspects of iconic formations are likely to interfere with the behaviour of these forms, although these interactions are less transparent than the semantic effects in compounds, which may allow for a slightly greater amount of phonological conditioning within this group. Finally, echo-pairs borrowed from other languages are unlikely to exhibit the patterns observed in Hungarian due to their foreign origin.

These arguments receive support from the proportions of labial forms within different morphological classes, which are shown in Figure 4.

While both compounds and iconic formations exhibit a higher proportion of forms with a labial behaviour than would be expected on the basis of the baseline

[8]The voicing of nasals is traditionally described as passive, mainly because of the fact that the maintenance of voicing in sonorants does not require any extra articulatory effort as opposed to obstruents (see e.g. Chomsky and Halle 1968). Passively voiced consonants do not normally exhibit a voicing contrast.

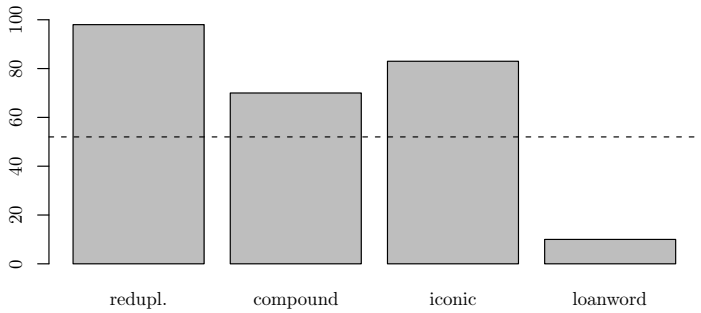

Figure 4: *The proportion of labials in classes of echo-pairs created through different morphological processes. The dashed line indicates the baseline, that is, the overall frequency of labial-initial forms in Hungarian on the basis of the Hungarian Webcorpus.*

frequency of labial-initial forms in Hungarian (indicated by the dashed line), these proportions are considerably lower than that observed among reduplicated echo-pairs.[9] Unsurprisingly, loanwords show a very different pattern: the proportion of labials is significantly lower than the baseline (for this reason, they will be excluded from the analyses presented in the rest of this section). These results suggest that the general pattern observed for reduplicated echo-pairs (i.e. the prevalence of labial forms) is present in compounds and iconic formations as well. The question is whether the more fine-grained patterns of phonological conditioning described above extend to these groups as well.

Since the combination of the MVDM and multidimensional scaling presented in the previous section does not yield any easily interpretable results for compounds and iconic formations, a different approach is taken. Instead of trying to isolate each of these tendencies within these groups, they are compared to the group of reduplicated forms in a more indirect manner. More specifically, an artificial learner is used to extract the tendencies among reduplicated forms and is then tested on compounds and iconic formations to see whether its knowledge of the phonological patterns in the former group can be successfully applied to the latter two groups. The rationale behind this strategy can best be understood through an analogy. Consider a native speaker of Danish trying to learn Norwegian and Hungarian. It is almost certain that they will be more successful at acquiring Norwegian than they will be at acquiring Hungarian (although, of course, these differences might diminish over

[9] All of the differences reported here are statistically significant at a level of $p < 0.01$ according to chi-squared tests with Yates' correction for continuity (standardly used for the comparison of proportions).

time). The reason for this is that many of the patterns in their native language can also be found in Norwegian but not in Hungarian. Similarly, an artificial learner trained on reduplicated forms will be more successful when tested on compounds or iconic formations if the latter two groups also contain some of the phonological patterns typical of reduplicated forms.

The artificial learner used for the purposes of this experiment is based on Nosofsky's (1986) Generalised Context Model (GCM). The basic principles of the GCM are as follows. Since the GCM is an exemplar-based model, the patterns in the training set are not learnt explicitly; instead, the GCM simply stores a feature representation of all the tokens in the training set along with their behaviour (the features in this case are the syllabic constituents in the third column of Table 2). These stored forms can all be accessed during the testing phase, when the model has to make predictions for forms whose behaviour is not known. The prediction of the behaviour of a given form is based on a token from the training set selected stochastically as a function of its similarity to the given form: the outcome of the prediction is simply the behaviour of this form.

The crucial step in the process described above is the selection of the token which serves as the basis of the prediction. As it has been noted above, the probability that a given token is selected is proportionate to its similarity to the form whose behaviour is not known. Similarity is calculated as a monotonically decreasing function of the distance between the two forms. The distance value is obtained by going through the feature representations of the two forms and keeping a tally of matching feature values. Features that are more important in the prediction can be also weighted, in which case they have a relatively higher influence on the distance value.[10] To give an example, the non-weighted distance between the forms *csicsa* [tʃ,i,=][tʃ,ɒ,=] 'kitsch' and *cica* [ts,i,=][ts,ɒ,=] 'cat' will be equal to the the distance between *csicsa* [tʃ,i,=][tʃ,ɒ,=] and *csacsi* [tʃ,ɒ,=][tʃ,i,=] 'donkey.*dim*' (the distance value is 2 in both cases). However, if the first and the second onset are more heavily weighted, *csicsa* and *csacsi* will be closer to each other than *csicsa* and *cica*. It should be noted that there are several important details about the similarity metric and the calculation of probabilities that have been omitted from the present discussion; the interested reader is referred to Nosofsky (1986) for a more detailed description of the GCM.[11]

[10] Thus, the first and the second onset were weighted more heavily than other features in the present experiment, since they were found to be relatively reliable predictors of the behaviour of echo-pairs in Section 3.1.

[11] Since the purpose of this experiment is to compare the performance of the same learner on different data sets, all the parameters in the model are kept constant. The specific values of the parameters are not particularly important for the present discussion, as they seem to affect the behaviour of the

Three simulations were run to test the performance of the artificial learner on the different data sets. The task in each simulation was to predict a behaviour for all the items in the test set. The training set was the same in each of these simulations: the set of reduplicated forms. The test set was varied between the simulations: the set of reduplicated forms was used in the first one, the set of compounds in the second one and the set of iconic forms in the third one. The motivation for testing the learner on the training set as well was to see how well it could learn the patterns within the group of reduplicated forms. Both the training set and the test set contained tokens rather than types, which means that the same forms often occurred several times within the same data set. Moreover, since the prediction mechanism in GCM is based on stochastic principles, tokens of the same word sometimes differed in their predicted behaviour.[12]

The results of the simulations are presented in Figure 5.

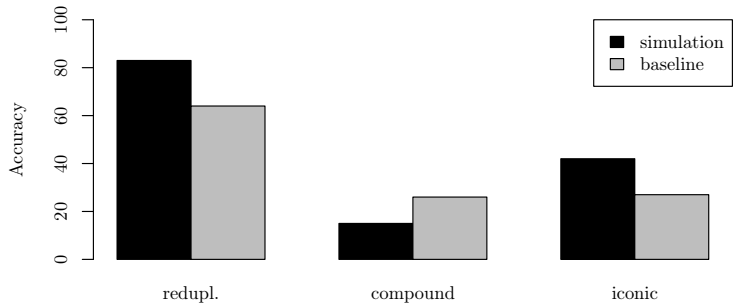

Figure 5: *The accuracy of the GCM compared to a baseline accuracy that could be achieved without using any of the fine-grained generalisations among reduplicated forms.*

The accuracy results for the simulations are obtained simply by dividing the number of correct predictions by the total number of forms in the test set. The baseline accuracy is defined as the accuracy that could be achieved by predicting the most frequent behaviour within the test set for all the items (i.e. the proportion of the most frequent behaviour within the test set). Unsurprisingly, the performance of the model is the highest when tested on reduplicated forms, surpassing both the baseline performance and the performance of the model for the other two data sets. While there can be no doubt that this is largely due to the fact that the training set

model similarly regardless of the data set it is tested on, and thus do not have a crucial effect on the comparative success of the model in the three different conditions.

[12]The artificial learner could not use different tokens of the given form to predict its behaviour, as this would have led to a nearly 100% success rate in the case where the training set and the test set were the same.

and the test set were identical, the difference between the baseline accuracy and the accuracy of the GCM corroborates the assumption that the phonological makeup of reduplicated forms can be used to predict their behaviour. The accuracy of the GCM is extremely low for compounds in comparison to both the baseline accuracy and the other two data sets. This suggests that the fine-grained phonological conditioning present among reduplicated forms is not found in this group. Finally, the performance of the GCM exceeds the baseline for iconic forms, but is markedly lower than its performance for reduplicated forms, which means that the phonological tendencies observed among reduplicated forms are present within this group, but have a weaker effect.

To sum up, it appears that both compounds and iconic forms show some of the general tendencies within the reduplicated group, but only iconic forms exhibit fine-grained phonological conditioning. This is compatible with the tentative claim made at the beginning of this section according to which semantic effects may interfere with the phonological patterns underlying echo-pair formation to a greater extent than sound symbolic considerations.

4 Discussion

This section presents a discussion of a number of theoretical points related to echo-pair formation. I begin by providing a brief summary of the observations made in the preceding sections. It will then be argued that the evidence about echo-pairs can best be explained if we assume that these forms are the result of the productive use of a product-oriented schema in the sense of Bybee (2001).

Let us review the evidence presented so far. There are several different word formation processes that can give rise to echo-pairs, including reduplication, compounding, the creation of iconic forms and borrowing. These processes appear to show a considerable degree of variation in the extent to which they are conventionalised. Thus, reduplication is geared specifically towards the creation of echo-pairs and exhibits a high level of systematicity in phonological terms. The creation of iconic forms is not restricted to echo-pairs and seems to be much less systematic as a means of echo-pair formation, although such forms also exhibit some of the phonological tendencies present in reduplicated forms. Compounding is an even more general morphological mechanism, which shows only the most general tendencies observed among reduplicated echo-pairs. Finally, loanwords do not seem to fit into any of the patterns discussed in the previous sections, which is likely to be a result of their foreign origin. It should also be noted that the phonological tendencies in Section 3 are all specific to echo-pair formation. The high proportion of labi-

als, the dissimilation of initial consonants and the pattern of long-distance voicing assimilation described for echo-pairs are not found anywhere else in Hungarian.

It is clear that models based on the application of symbolic morphological rules (see e.g. Katamba 1993) are incapable of capturing the regularities (or, rather, irregularities) described above. Several separate morphological rules would have to be posited to account for the different morphological processes described above, despite the fact that echo-pair formation clearly exhibits a number of shared properties that are seen in all of these classes. It is also difficult to see how non-conventionalised processes such as the creation of echo-pairs through compounding could be captured in any insightful way in such a framework. Moreover, the variability inherent in echo-pair formation also cannot be straightforwardly represented in a model based on categorical morphological rules.

Therefore, I believe that the evidence presented above calls for a different approach. I propose that echo-pair formation can best be explained as the result of the productive application of a morphological schema. A *schema* can be defined as a collection of phonological generalisations describing a set of forms sharing a similar set of functions (cf. Bybee and Slobin 1982; Bybee 2001). Whether such schemas are represented explicitly in the form of phonological statements or emerge from patterns among stored forms is immaterial to the present discussion: echo-pair formation is compatible with both views and the evidence reviewed above does not favour either of them. Some of the phonological generalisations associated with echo-pair formation are as follows: altered repetition of the same phonological sequence, labial-initial second component, dissimilation between the initial consonants of the two components and voicing assimilation between the second onset and the initial consonant of the second component. The set of shared functions include diminution, hypocorism and connotations of 'playfulness' more generally. Since the phonological tendencies above are directly associated with the semantic functions of echo-pairs, it is not surprising that they are not found elsewhere in the language.

Echo-pairs can best be described through a product-oriented schema (see Bybee 2001), which does not prescribe the way a particular morphological construction is assembled. A form that conforms to the phonological generalisations pertaining to echo-pairs will automatically be associated with the range of functions typical of echo-pairs regardless of the morphological pathways through which it is created (thus, both *csillog-villog* and *cica-mica* have connotations of playfulness and informality despite the fact that one of them is created through compounding and the other through reduplication). Product-oriented schemas can be contrasted with source-oriented schemas, which also determine the way a given form is constructed. For instance, a form such as *lens* in English will not normally be interpreted as

plural due to the absence of a corresponding singular form *len* and the implied impossibility of *lens* being derived from a base form.

Importantly, product-oriented schemas do not have to be completely categorical: a form that only satisfies a subset of the phonological generalisations embodied in a schema can still be associated with the relevant meanings. Thus, *cica-mica* satisfies all the phonological constraints on echo-pairs, while *Tapsi-Hapsi* 'Bugs Bunny' only the most general of them; however, they both have playful, diminutive associations. It could be further assumed that forms that fit more of the phonological patterns seen among echo-pairs will be associated with more prototypical functions within this group. Unfortunately, the present data set does not provide any means of testing this hypothesis.

In conclusion, morphological schemata provide a straightforward way of capturing the phonological and morphological properties of echo-pairs in Hungarian. We can account for the fact that echo-pairs can be created through several different morphological pathways through the assumption that this phenomenon is based on a product-oriented schema. The variation in the behaviour of echo-pairs can be interpreted as a consequence of the non-categorical nature of schemata. Finally, the observation that the phonological tendencies in echo-pair formation are not found elsewhere in Hungarian follows from the inclusion of these tendencies in the morphological schema describing echo-pairs.

5 Conclusion

In this paper, I have presented a phonological and morphological analysis of echo-pair formation in Hungarian. It has been shown that even such a highly variable and unconventional pattern can be investigated systematically through the use of modern computational and statistical methods. Thus, a corpus search has been used to compile an extensive data set consisting of echo-pairs, which has been analysed through a variety of computational techniques. Several phonological tendencies have been identified, including an unusually high proportion of labial forms, a pattern of dissimilation between the initial consonants of the two components of echo-pairs and a pattern of long-distance voicing assimilation. It has also been demonstrated that these tendencies can be found in all three classes of echo-pairs created language-internally, although their strength has been found to vary across these classes. Finally, I argued that the evidence related to echo-pairs supports a non-symbolic approach to morphological patterns, where the creation of novel word-forms is potentially guided by product-oriented schemata.

Bibliography

Apor, D. (1906). *Az ikerszók* [Echo words]. Budapest: Engel S. Zsigmond.

Bell, A. (1984). Language style as audience design. *Language in Society 13*, 145–204.

Blevins, J. and A. Garrett (2004). The evolution of metathesis. In B. Hayes, R. Kirchner, and D. Steriade (eds.) *Phonetically driven phonology*, 117–156. Cambridge: Cambridge University Press.

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

Bybee, J. L. and D. I. Slobin (1982). Rules and schemas in the development and use of the english past tense. *Language 58*, 265–289.

Chomsky, N. and M. Halle (1968). *The sound pattern of English*. New York, Evanston, London: Harper & Row.

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

Daelemans, W., J. Zavrel, K. van der Sloot, and 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, and V. Trón (2004). Creating open language resources for Hungarian. In *Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004)*, pp. 203–210. Lisbon: European Language Resources Association.

Hansson, G. Ó. (2004). Long-distance voicing agreement: An evolutionary perspective. In M. Ettlinger, N. Fleischer, and M. Park-Doob (eds.) *Proceedings of the 30th Annual Meeting of the Berkeley Linguistics Society*, 130–141. Berkeley: Berkeley Linguistics Society.

Jakobson, R. (1965). Quest for the essence of language. *Diogenes 51*, 21–37.

Katamba, F. (1993). *Morphology*. New York: St. Martin's Press.

Keane, E. (2001). *Echo words in Tamil*. Ph. D. thesis, Oxford University.

Labov, W. (1994). *Principles of Linguistic Change, vol. I: Internal Factors*. Oxford: Blackwell.

Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General 115*, 39–57.

Prince, A. and P. Smolensky (1993). Optimality Theory: Constraint interaction in generative grammar. Manuscript, Rutgers University and University of Colorado at Boulder. Available at ROA.

Simonyi, Z. (1907). *Die ungarische Sprache, Geschichte und Charakteristik*. Strassburg: K. J. Trübner.

Siptár, P. and M. Törkenczy (2000). *The phonology of Hungarian*. Oxford: Oxford University Press.

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

Szikszainé Nagy, I. (1993). *Az ikerítés helye, szerepe, szabályszerűségei a magyar nyelvben* [The place, role and rules of reduplication in Hungarian]. Budapest: Magyar Nyelvtudományi Társaság Kiadványai, vol. 197.

Thun, N. (1963). *Reduplicative words in English*. Uppsala: Carl Bloms Boktryckeri.