Eötvös Loránd Tudományegyetem
Bölcsészettudományi Kar

Hungarian Phonology and Morphology

—

Discord in the Possessive Allomorphy of Hungarian

Magyar Fonológia és Morfológia

—

A Magyar Birtokos Allomorfia Disszonanciája

Master's Thesis / Szakdolgozat

Supervisor / Témavezető: Rebrus Péter
Author / Szerző: Rácz Péter

Theoretical Linguistics MA
Elméleti Nyelvészet szak

Budapest, 2010

CERTIFICATE OF RESEARCH

By my signature below, I certify that my ELTE M.A. thesis, entitled Hungarian Phonology and Morphology – Discord in the Possessive Allomorphy of Hungarian, is entirely the result of my own work, and that no degree has previously been conferred upon me for this work. In my thesis I have cited all the sources (printed, electronic or oral) I have used faithfully and have always indicated their origin.

Date: 15 April 2010

Nyilatkozat

Alulírott, Rácz Péter, kijelentem és aláírásommal igazolom, hogy szakdolgozatom, melynek címe 'Magyar Fonológia és Morfológia – A Magyar Birtokos Allomorfia Disszonanciája', saját munkám eredménye. A szakdolgozatot korábban diplomamunkaként nem használtam fel. A felhasznált irodalmat korrekt módon kezeltem, eredetét mindig feltüntettem.

2010 április 15

Ich singe keine Melodie, Ich singe eins zwei drei vier.
The Jeans Team

Acknowledgements

I would like to thank my fairy godmother, Péter Rebrus, for his wise sentiments, Péter Siptár for the astute remarks, as well as my excellent Klassenkameraden, Márton Sóskuthy and Dániel Szeredi (in alphabetic order) for all the emails. Further thanks are due to Christian Uffmann, Koen Sebregts, and Michael Schäfer for the suggestions of großartiger deutscher Elektropop to enjoy during writing. I am grateful to Kinga Gárdai for her help in humdrum bureaucratic matters, Ágnes Füle for defending my credit transfer documents with all the savage ferocity of a she-tiger at bay, Bernd Kortmann for his patience with my freelance career in morphology, and last, but not least, Flóra Horváth, for everything.

Contents

1 Introduction

This paper tries to tackle possessive allomorphy in Hungarian[1]. I will argue that the behaviour of the 3SG-POSS form cannot be explained only on phonological grounds, and that one of its main shaping factors is analogy, interpreted as the effect of similar forms stored in the lexicon. I will also claim that we have to assume an exemplar-based representation in order to fully describe 3SG-POSS variation and the possible factors that play a role in it. While the 3PL-POSS also shows variation, the scope of this paper is exclusively the singular.

The fact that this variation is unnatural or phonologically sub-optimal proves to be a serious problem for approaches that try to analyse allomorph selection as exclusively grounded in phonological optimisation. Hungarian possessive is not the only example of phonological 'discord' in allomorph distribution. The novelty of this analysis is that it does not only point out the difficulties of grounding it in phonological optimisation, but it also tries to provide an alternative explanation.

This explanation builds on a very general notion of analogy as the effect of similar forms in the lexicon. Such analogical levelling processes can be regarded to take place during language change, and their mechanics require us to eschew minimal memory models of lexical storage, as lexical frequency has a prominent role in these levellings, and it finds no place in a redundancy-free lexicon.

What this paper offers, then, is a satisfactory explanation of the patterns observed in the variation of the possessive, specifically, the 3SG-POSS. It also provides evidence for a view of lexical storage which is detailed and organised based on general similarity, and stresses the importance of diachronic explanations in linguistics. The two aspects are necessarily tangled up, as it is the rich memory model in which the analogy-based diachronic explanation can be assessed. In this sense I follow a line of research put forward by, among others, Bybee (2001) and Blevins (2004).

The paper is organised as follows: in section 2, we survey the workings of the Hungarian possessive in general, then, in section 3, we go into the details of the variation displayed by the 3SG-POSS. Section 4 discusses it in the context of naturalness and allomorphy, and section 5 provides a diachronic, analogical

[1]In Hungarian, the letters *a,e,á,é* roughly correspond to the IPA symbols ɒ,ɛ,aː,eː, respectively. Phonetic transcriptions are only provided when relevant.

account for it. Section 6 discusses the results of a partial simulation of the allomorph distributions using the Tilburg Memory-based Learner (Daelemans et al., 2007). Section 7 gives brief conclusions.

2 The Hungarian possessive

This section surveys the basic allomorph distributions of the Hungarian possessive. The loci of interest are the 3SG and 3PL forms, and the rest of the paper concentrates exclusively on the 3SG-POSS. Nonetheless, the whole of possessive formation is crucial to our analysis. This discussion draws heavily on Papp (1975) and Kiefer (1985), as well as on Kiefer (2000).

2.1 Allomorphy

The possessive is suffixed to the possessee and marks the person and number of the possessor. The basic paradigm can be seen in table 1 with a vowel-final stem, a consonant-final back vowel stem, and a consonant-final front vowel stem. The 3SG-POSS is highlighted.

	kapu	asztal	szék
1SG	kapu*m*	asztal*om*	szék*em*
2SG	kapu*d*	asztal*od*	szék*ed*
3SG	kapu***ja***	asztal***a***	szék***e***
1PL	kapu*nk*	asztal*unk*	szék*ünk*
2PL	kapu*tok*	asztal*otok*	szék*etek*
3PL	kapu*juk*	asztal*uk*	szék*ük*
Gloss	'gate'	'table'	'chair'

Table 1: The possessive paradigm

The endings have a typical set-up of a linking vowel (depending on the environment) and a consonant-initial suffix, except for the 3SG and 3PL forms. The consonant-initial suffix forms take a linking vowel with consonant-final stems, the vowel agreeing with the backness of the stem vowel. The vowel-initial forms have the epenthetic glide *j* when following vowel-final stems (*j* is the only glide in Hungarian).

If the stem ends in a low mid vowel (ɛ or ɒ), the vowel gets lengthened before suffixes: *alma* [ɒlmɒ] 'apple'; *almája* [ɒlmaːjɒ] 'apple-3SG-POSS; *körte* [kørtɛ] 'pear'; *körtéje* [kørteːje] 'pear-3SG-POSS.

The linking vowel also agrees with the stem vowel in roundness if the stem vowel is a front vowel (Hungarian has no unrounded back vowels). Furthermore, while the linking vowel tends to be a mid vowel (with all suffixes, except for the 1PL-POSS), some lexically specified stems take a low linking vowel instead. These are the so-called 'lowering stems'. So, instead of **tüzöm*, we find *tüzem*[2]. Rounding harmony and lowering are exemplified in table 2.

	ezüst	tűz
1SG	ezüst*öm*	tüz*em*
2SG	ezüst*öd*	tüz*ed*
3sg	ezüst*je*	tüz*e*
1PL	ezüst*ünk*	tüz*ünk*
2PL	ezüst*ötök*	tüz*etek*
3pl	ezüst*jük*	tüz*ük*
Gloss	'silver'	'fire'

Table 2: The possessive paradigm

In the case of 'silver-3SG-POSS', the form takes a glide-initial suffix, despite the fact that the stem itself ends in a consonant (*ezüst-je*). This is because the glide does appear after C-final stems, but not categorically. The situation is as follows:

- The 3SG-POSS suffix has four possible allomorphs:
  - Back and front vowel forms without an initial glide: *-a, -e* (*-A*)
  - Back and front vowel forms with an initial glide: *-ja, -je* (*-jA*)
- V-final stems always have *-jA*:
  - cf. *kapu-ja* 'gate-3SG-POSS', *eskü-je* 'oath-3SG-POSS'

[2]Compare with *bűzöm* 'stink-1SG-POSS'.

  - This distribution is categorical.

- Sibilant/palatal C-final stems always have *-A*:
  - Cf. *gáz-a* 'gas-3SG-POSS', *té[ɲ]-e* 'fact-3SG-POSS'
  - This distribution is also categorical with minor counterexamples (*a[ɲː]-a* (any+ja) 'mother-3SG-POSS', *ci[s]-je* 'C#-3SG-POSS').

- Other C-final stems vary between *-A* and *-jA*:
  - cf. *nép-e* 'people', *nap-ja* 'sun'

In sum, the 3SG-POSS is a harmonising suffix, and it has a glide-initial form *-ja, -je* (*-jA*) after vowel-final stems and a form without an initial glide, *-a, -e* (*-A*), after stems ending in palatal or sibilant consonants. The appearance of the glide varies elsewhere. (The 3SG-POSS will be referred to as *-(j)A*.) To give an example of a descriptive account, Rebrus (2000) analyses this suffix as two morphemes, *-j-* and *-A*, as these can appear independently, depending on person/number and stem ending. For instance, none of them appears in forms like *hajó-i* 'ships-3SG-POSS', *ház-uk* 'house-3PL-POSS'. Both appear in e.g. *zokni-j-a* 'sock-3SG-POSS, while we find only *-A* in *ház-a* 'house-3SG-POSS. While such an account impeccably captures the possible distributions, it fails to explain the alternation between *-jA* and *A*. The variation between *-jA* and *-A* is surveyed in detail in the next section. First, we briefly look at the history and present-day status of this allomorphy.

2.2 The emergence and grounding of *-jA*

In Old Hungarian, the possessive marker did not have a glide-initial allomorph. It developed later, and occurred first with back vowel-final stems. By the 17th century, it showed both with back and front vowel-final stems, but still with a preference for the latter (Papp, 1975). Later on, it began to occur after consonant-final stems as well. According to contemporary grammars of Hungarian, a glide-initial allomorph was possible after sibilant- and palatal-final stems in the late 18th and early 19th century (Kiefer, 1985). At this stage, the language allowed forms like *olaj-ja* [olɒjːɒ] 'oil-3SG-POSS', *sas-ja* [ʃɒʃːɒ] 'eagle-3SG-POSS', or *bárány-ja* [baːraːɲːɒ] 'sheep-3SG-POSS'. By the end of the 19th century, however, this was no longer allowed in Standard Hungarian, and was considered dialectal in later grammars. While such a

diachronic change is possible, this one is conspicuous, as it has no phonological grounding. We will explore this problem later.

In the context of the change in possessive allomorphy, it is important to note that novel loanwords and proper nouns have a very strong tendency to take the *-jA* form, such as in *fájl-ja* 'file-3SG-POSS', *Napóleon-ja* 'Napoleon-3SG-POSS' – except when they end in sibilants or palatals. Data from child language also show a preference of the *-jA* form over the *-A* form, despite the fact that, as we will later see, the glide-initial form occurs only with a minority of consonant-final stems.

It is easy to find phonological grounding for the categorical distributions of *-(j)A*. The obligatory *-jA* after vowel-final stems can be explained by donating a hiatus-filling function to the glide. This is further corroborated by the history of the possessive, as the glide-initial allomorph first showed after vowel-final stems.

The categorical avoidance of the glide-initial allomorph after sibilant and palatal consonants could be based on articulatory/perceptual grounds. One could argue that the sibilant/palatal consonant plus glide sequences are harder to produce and perceive. Indeed, there are several examples of the avoidance of consecutive identical features in the literature (McCarthy, 1986). Besides, the avoidance of these sequences can be observed in Hungarian morphology, such as in the verbal paradigm.

The verbal imperative/subjunctive endings are *j*-initial. If the stem ends in a sibilant or a palatal, these initial glides are assimilated. Table 3 illustrates this with the singular paradigm of the imperative with stems ending in a velar stop, a palatal stop, and a sibilant fricative, respectively. In the latter two cases, the imperative glide assimilates to the stem consonant.

Inf	rak-ni	rogy-ni	húz-ni
1SG	rak*[ç]ak*	ro*[ɟ:]ak*	hú*[z:]ak*
2SG	rak*[ç]*	ro*[ɟ:]*	hú*[z:]*
3SG	rak*[ç]on*	ro*[ɟ:]on*	hú*[z:]on*

Table 3: Glide assimilation in the verbal paradigm

The definite ending in the 3SG and all the plural forms behaves similarly: *rak* 'put-3SG-INDEF vs. *rakja* 'put-3SG-DEF'; *hagy* 'leave-3SG-INDEF' vs. *ha[ɟ]a* 'leave-3SG-DEF'.

This hints at a tendency to avoid palatal/sibilant consonant plus *j* clusters in Hungarian morphology. We could say that while the verbal paradigm opts for assimilation to avoid the illicit clusters, the nominal possessive paradigms erases (or excludes) the glide altogether. It is an important question why the verbal paradigm goes for assimilation and the nominal for deletion, and we will return to it later in this paper.

In this section we surveyed the general properties and the categorical distributional patterns of the possessive. The next section turns to the variation it displays and the observable patterns in this variation. As indicated above, we will concentrate on the 3SG-POSS, as it displays most of the non-categorical behaviour in possessive allomorphy.

3 Variation

The previous section discussed the categorical distributions of *-(j)A* allomorphy, such as that the glide is obligatory after vowels and categorically absent after sonorants and palatals. This section takes stock of the variation it displays. We will find patterns requiring further analysis as there are no satisfactory phonological explanations for them. The range of variation surveyed here encompasses the following aspects:

1. Vowel quality (back versus front vowel stems)
2. Final consonant quality (place and manner of articulation)
3. Final consonant quantity (light syllables versus heavy syllables versus consonant clusters)

The data are drawn from the Hungarian Webcorpus (Halácsy et al., 2004), a corpus of 1.48 billion words from 18 million pages downloaded from the .hu Internet domain, which gives the best representation of written language, and is the most faithful corpus of present-day Hungarian. Of course, it has its limitations: since its sources are written, the register is slightly more conservative, which severely curbs its ability to represent spoken language, as we will see in the case of individual word variation with *-(j)A* allomorphy. In the following tables in this section, numbers represent thousands, and, unless otherwise noted, indicate token frequency. Type frequencies are also given in thousands.

3.1 Tokens and types

The variation patterns below are given in *tokens*. There are several reasons to believe that token ratios are more important than type ratios in determining the possible factors affecting allomorph variation. Mainly, as these factors will take the shape of analogical relations, frequency will be very relevant for them. Bybee (2001) claims that type frequency is the relevant factor in analogical levelling inasmuch as the pattern with more types in it will exert an influence on the pattern with the minority of types. Past tense formation in English is a classical example. Here, the dental suffix class, to which the majority of verbs belong, slowly devours the irregular strong verbs. Token frequency is only relevant as it can protect a verb form from levelling, so that verb types with high token frequency will retain their irregular strong past tense forms. Verbs of high token and low type frequency can only exert influence on very similar novel items, resulting in the limited productivity of the strong verb set (Rumelhart & McClelland, 1987).

Then again, Bybee's direction of levelling cannot always be witnessed (Labov, 2006). Besides, we have other reasons to think that token-based calculations will prove more satisfactory in our case. Firstly, Rebrus & Törkenczy (2008) and Zsuzsa Kertész & Kálmán (2008) suggest that analogy based on similarity of form is generally token-based, though type ratios also play a role. I will argue that such an analogical relation is active in *-(j)A* allomorphy.

Secondly, in this case, token ratios usually correspond to similar type ratios. When there is a difference between the token and type frequencies of various classes, the type frequencies are distorting factors, since there are a large number of types with only a few tokens, exerting minimal influence on the other types.

One example of this is that the corpus contains far more back vowel types than front vowel types, but this is not true for the tokens. As we will see, certain *-(j)A* allomorphs seem to prefer back stems, but the weight of types cannot be a reason for this, as this skewed distribution is absent with other allomorphs, or indeed, other suffixes. Table 4 gives two examples of the token/type asymmetry of front/back stems, illustrated with the 3SG possessive (*-(j)A*) and dative (*-nAk*) suffixes. While there are substantially more back vowel types in both cases, this distinction evens out when we look at the token ratios.

As can be seen, while the token ratios support our intuition that a language

Suffix	Token	Type	Token ratio	Type ratio
-(j)a	5694	18	0.50	0.62
-(j)e	5672	11	0.50	0.38
Sum	11366	29	1	1
-nak	1858	22	0.48	0.65
-nek	2001	12	0.52	0.35
Sum	3859	34	1	1

Table 4: The *3sg* possessive and dative with front and back vowel stems

with vowel harmony should contain approximately the same amount of forms from the possible vowel sets, there seem to be an excess of back vowel types. These types have very few tokens and do not play a role in the processes outlined below. Apart from such examples, token ratios are usually faithful to type ratios. Based on all this, it could be argued that if there is a discrepancy between type and token ratios, tokens should be preferred to types.

3.2 Allomorph distributions

While the categorical distributions of the 3SG possessive have phonological grounding, the patterns of variable distributions are harder to explain. That is to say, while the initial glide of the *-jA* form of the possessive can be argued to fulfil a function as a hiatus filler after vowel-final stems, it is less clear why it is preferred to an extent also with consonant-final stems after back vowels or consonant clusters (which it is). We list some of these unnatural patterns below.

The first and strongest is the preference of back vowel stems for the glide-initial allomorph (*-jA*). As it can be seen in table 5, 2 per cent of front vowel consonant-final stems select the *-jA* form in the corpus, while 16 per cent of back vowel consonant-final stems do. The numbers indicate token frequency.

What follows is that 87 per cent of the *-jA* forms are back vowel forms. This asymmetry is atypical for a harmonising suffix, which can be seen in table 4 above: both the dative and the total of the possessive *-(j)A* have an equal amount of front and back vowel tokens. It is also important to note that this asymmetry cannot be explained on phonological grounds. If we regard

Stem vowel	*-jA*	*-A*	Sum	j/all ratio
Back V	807	4105	4912	**0.16**
Front V	117	5066	5183	0.02
Sum	924	9171	10095	0.09

Table 5: Back and front vowel (consonant-final) stems with *-(j)A*

the glide a hiatus filler, it should not occur after consonants at all. In any case, it should not be sensitive to the vowel quality of the stem. Therefore, this is an asymmetry that requires further, non-phonological explanations.

Stem ending is the second respect in which the 3SG-POSS shows skewed distributions. As table 6 shows, the *-jA* form (the one with the glide) prefers stems ending in consonant clusters to stems ending in a single consonant. Furthermore, short vowels are preferred to long vowels (Hungarian has no diphthongs).

Stem ending	*-jA*	*-A*	Sum	j/all ratio
VCC	287	559	846	**0.34**
VVC	317	6312	6629	0.05
VC	320	2300	2620	0.12
Sum	924	9171	10095	0.09

Table 6: Stem endings and *-(j)A*

This pattern also deviates from the norm. At most, we would expect a negative correlation between the number of stem-final consonants and the presence of the glide, as an intervocalic CCC cluster is more marked than a CC one. What we find, though, is that 67 per cent of the *-jA* forms have a stem ending in a consonant cluster. Even if we admit that the glide is not a simple hiatus filler, since it appears after consonants, it should not be sensitive to the amount of preceding consonants. Indeed, no similar alternations in Hungarian are sensitive to such information. The comparison of VC final stems and VCC final stems would hint at the role of weight here, but VVC final stems, which should count as heavy as VCC ones[3], prefer the *-jA* form

[3]If not more, as the second consonant of the cluster could be re-syllabified before

less.

Final C	coronal stop	labial stop	velar stop
-jA	350	185	150
-A	1038	125	1706
Sum	1388	311	1856

Table 7: Stem endings and *-(j)A*: final stops

The third interesting pattern is that the preference for *-jA* forms shows a correlation with the quality of the stem-final consonant as well (cf. table 7). The ratio of *-jA* versus all 3SG-POSS (*-(j)A*) is 0.25 with final coronal stops, 0.59 with labial stops, and 0.08 with velar stops – it is 0.00 with sibilants and palatals. (Recall that the total average is 0.09 (*-jA* vs. *-(j)A*).) This is only important as it shows that there are very complicated sub-regularities in the distribution of the 3SG-POSS, including an extent of sensitivity to the place of articulation of the word-final consonant. We can observe sensitivity to the manner of articulation as well: for instance, while 59 per cent of forms with labial stop stem endings have the *-jA* forms, only 0.1 per cent of forms with labial fricative endings do.

This could just as well be a side effect: frequent nominal derivative suffixes end in *k,g,t*, and these avoid *-jA*, while it is obligatory after comparatives, which end in *b* (cf. next section). The detailed analysis of the interaction between possessive variation and derivative morphological classes is beyond the scope of this paper.

3.3 Further subregularities

The preference for the glide form *-jA* does not only correlate with phonological properties. Further determining factors include morpho-phonological, semantic, and lexical ones. A few main non-phonological factors are listed below.

- Morpho-phonological

assigning syllable weight.

  - Lowering stems, that is, stems taking a low linking vowel instead of a mid one, tend to avoid the *-jA* form. An example for a lowering stem is *vár* 'castle', cf. *vár*-**a** 3SG-POSS, *vár*-**a***m* 1SG-POSS, *vár*-**a***k* PL, *vár*-**a***t* ACC, etc.

- Morphological

  - Particular derivational suffixes always go for *-A*, such as *-sÁg, -At, -nOk*, e.g. *visszássága* 'perversity-3SG-POSS', *csillagászata*, 'astronomy-3SG-POSS'. The frequency of these suffixes may explain the correlations between stem-final consonant quality and possessive allomorph selection.

- Semantic

  - Inalienable possessions tend to have the *-A* form, whereas alienable possessions have the *-jA* form. An example from Kiefer (1985) is *ablak*-a/*ablak*-ja 'window-3SG-POSS'. The former form would refer to the window of a building, whereas the second to the window of a stock of windows in a warehouse or the collection of a window collector.

- Lexical

  - Frequent words can behave individually. First, frequent words are the ones that are most likely to display word-level variation. One example is *német* 'German'. Most native speakers would accept both the forms *némete/németje* for the 3SG-POSS – though this can also have determining semantic contexts. Second, with particular high frequency forms the two allomorphs can get lexicalised, as with *kar* 'hand/choir', where the form with the meaning 'hand' takes *-ja*, while the one with the meaning 'choir' takes *-a*.

Variation does not only entail that phonologically very similar words prefer different possessive allomorphs, as in the case of *napja* 'sun-3SG-POSS'/*képe* 'picture-3SG-POSS', but also that individual words can take different suffixes, as in the case of *német* above. This type of variation is somewhat tricky to assess based on the Hungarian Webcorpus, as people seem to be more conservative and consistent in writing, avoiding word-level variation. This is suggested by the fact that the forms sounding natural for a native speaker

(suggesting large variation) show up in conspicuously small amounts in the corpus. In addition, the part-of-speech tagging of a number of forms that are prone to show variation (like the nouns also functioning as adjectives, such as *német* 'German') is intricate, as some are tagged as adjectives while others as nouns, rendering general comparison very difficult.

In order to gain data on word-level variation, the following method was employed: if a word occurs in both forms (i.e both with *-jA* and *-A*) it was categorised as 'variable'. This builds on the assumption that if a word in the corpus shows variation at all, it will show much larger variation in the spoken language. The co-variable investigated was frequency. Table 8 shows the extent of variation, as defined above, in logarithmic steps of token frequency.

Frequency in corpus	Variable types
1-699	1%
700-6999	10%
7000-69999	19%
70000+	24%

Table 8: Variation and frequency

Since the total amount of the relevant 3SG-POSS forms in the corpus is around seven million (these have stems ending in a non-palatal, non-sibilant consonant), a tenfold increase based on this number was set to distinguish the frequency classes. The upper cut was seventy thousand. The variation in these classes was determined by taking a random sample of a thousand types and by looking at how many have both forms in the corpus. This sample was cut to two hundred with the smallest group, simply because there was an insufficient number of types with a high enough token frequency to permit a thousand-strong sample.

The results clearly show that word-level variability correlates with token frequency. The more frequent the type is, the more likely it is that it occurs with both the *-jA* and the *-A* form in the corpus. This is another pattern that requires explanation.

4 Naturalness and Morphology

The variation observed in the possessive allomorphy, specifically, the 3SG-POSS, touches upon the question of naturalness in morphology. It has been a widely shared assumption that allomorph selection creates phonologically less marked or more natural structures, and that the distribution of allomorphs can be explained on grounds of phonological optimisation.

4.1 Allomorph selection as optimisation

The concept that allomorph selection is driven by phonological optimisation is made explicit in Prince & Smolensky (1993). One of their examples is Tagalog infixation. They observe that the positioning of the infix always results in less marked syllable structure, opting for CV.CV instead of VC.CV. Table 9 gives an example.

um+tawag	→ t-*um*-awag	'call, pf., actor trigger'
um+abot	→ *um*-abot	'reach for, pf., actor trigger'

Table 9: Tagalog Prefixal Infixation (Prince & Smolensky, 1993, p.34)

In the first case, they argue, the position of the infix results in the optimal CV.CV structure, which violates neither the constraint for obligatory onsets, ONS, nor the one against codas, NOCODA. In the second case, this would only be possible by pushing the infix further into the word, which would violate faithfulness constraints on word shape.

This analysis determines the position of the suffix by the means of constraint interaction. The idea that infixation in Tagalog favours less marked syllable structure is not new. Prince and Smolensky point out that it was already suggested by Anderson (1972), who used a complicated rule-based analysis to account for it.

Word formation favouring less marked structures is also the central tenet of Natural Morphology (NM) (Dressler et al., 1987; Dressler, 1999). NM approaches morphological processes from a functional-pragmatic point of view: a morphological structure is optimal or maximally natural if it is iconic, uniform, and transparent. Iconicity here means that a symbol referring to a more marked semantic structure is also more marked. For example, we can

say that plural is more marked than singular, so a language using a more complex structure for the plural satisfies this condition: the plural of English *table* is *tables*, a more complex structure.

Uniformity means that a function is expressed identically in all cases. In this sense, English fares better at marking plural than German, since the vast majority of English nouns take a *-s* suffix, whereas German has more ways to express the plural, including the suffixes *-s, -e, -en, -er*, and/or Umlaut. Finally, transparency means that the forms responsible for different functions are distinguishable themselves. In this sense, a form like *bicycle-s* (sg. *bicycle*) is more natural than a form like *Fahrräd-er* (sg. *Fahrrad*).

The prediction of NM is that most morphological processes will tend to be natural, and that language change also proceeds from unnatural to natural. In addition, natural classes will be the productive ones. To take an example from Kiefer (1985), the English plural *boy–boy-s* is maximally iconic (and, hence, natural), *goose–geese* is minimally iconic, and *sheep-sheep* is noniconic, and the degree of naturalness corresponds to both the direction of change and productivity. The *-s* suffix set is taking over the English plurals, and it is the only productive one.

The concept of phonologically-driven allomorph selection does not only work with the positioning of a single suffix, or the selection of a surface allomorph for an underlying form. It can be extended to cases of suppletion, where the forms expressing a single function share no underlying form, but their distribution is still phonologically determined. Carstairs-McCarthy (1988) gives the example of Hungarian 2SG-INDEF verbal suffixation, where the two possible allomorphs, *-Vl* and *-sz*, give no reason to posit a single underlier but occur in a phonologically predictable manner: the former after sibilants, and the latter after everything else:

- *ráz* 'shake-3SG-INDEF' – *rázol* 'shake-2SG-INDEF'
- *kap* 'get-3SG-INDEF' – *kap[s]* 'get-2SG-INDEF'
  - V-2SG-INDEF → [V] +[-ul] / [sibilant]_
  - V-2SG-INDEF → [V] +[-s] / elsewhere

Rubach & Booij (2001) build on the notion of optimal allomorph selection without a single underlying representation in their analysis of Polish iotation. They claim that this process is much easier to analyse in terms of different

allomorphs than by the use of a complicated palatalisation rule. The role of phonology, then, is not to derive the observable output forms, but rather to take care of their distribution. In their view, 'the allomorphs are idiosyncratic but their distribution is not' (Rubach & Booij, 2001, p.36).

Allomorphy can entail non-categorical behaviour as well. Anttila (2002, 2007) takes several phenomena in Finnish under scrutiny to reach the conclusion that variation can be described in terms of partially ordered constraints. Specifically, morphological conditioning can emerge if phonological constraints underdetermine the input. If the input is generally underdetermined, we observe variation.

The tendencies displayed by morphological variation, then, are handled similarly to those displayed by phonological variation. Optimality Theory has the proper toolkit for this, in the form of stochastically ordered constraints (Boersma, 1998). Hayes & Cziráky Londe (2006) and Hayes et al. (2009) use Stochastic OT to model variation, along with its potential causes, in Hungarian vowel harmony. Hayes & Cziráky Londe (2006) give an account on the variation of the dative suffix after vowels like the vowel *e*. These vowels are swifting from being neutral to patterning with the front vowels in vowel harmony. The weighting of constraints allows the model to reproduce the observed variation patterns faithfully.

Hayes et al. (2009) dwell on the issue of the possible factors influencing variation. They reach the conclusion that if the allomorph distribution is unstable, the patterns of variation can be sensitive to various phonological factors, such as the shape of certain word classes, distance, number of vowels, etc. Such factors play a role both in the selection of a front or a back vowel allomorph after the unstable vowel *e* in Hungarian, and in the selection of the glide-initial 3SG-POSS after consonant-final stems.

In sum, there is a strong torrent in the literature representing the view that allomorph selection always follows from (phonetically based) phonological optimisation. This view can handle cases where the position of a single morpheme has to be determined, as in Tagalog. It can also account for the distribution of several allomorphs to which we can posit a single underlying form, and can also make predictions on the behaviour of suppletive forms, still in the frame of phonological optimality. The two questions that readily lend themselves to consideration are (i) whether phonological optimality can be held responsible in all cases of allomorph selection and (ii) specifically, whether it can account for the behaviour of the 3SG-POSS in Hungarian.

4.2 Problems with the optimisation approach

The theory that allomorph selection follows from phonological optimisation faces difficulties with two types of phenomena. These are when the distribution of affixes does not seem to satisfy markedness or naturalness requirements (cases of phonological 'discord') and when there is no output for a specific morphological function at all (cases of paradigmatic gaps).

Examples of paradigmatic gaps include Hungarian (Rebrus & Törkenczy, 2010) and Norwegian (Rice, 2005). In Norwegian, the imperative is formed by a truncation of the infinitive, which usually has the stem plus a final *e*. If the form resulting from the truncation violates the syllable structure constraints of the language (it has a cluster of rising sonority word-finally), it will be ill-formed, and we get no output for the imperative form. This is shown in table 10. While the truncated form of *skrive* 'write-INF', *skriv*, is a well-formed output, the truncated form of *cykle* 'cycle-INF', **cykl*, is not.

skrive 'write-INF'		*skriv*!
opne 'open-INF'	*open* open (A)	**opn*!
cykle 'cycle-INF'	*cykel* bicycle (N)	**cykl*!

Table 10: Norwegian imperatives

It is important to note here that it is possible to repair illicit word-final clusters by epenthesis in Norwegian, which can be seen on the adjectival and nominal forms of the defective verbs above. This, however, is impossible with the imperatives. Such paradigmatic gaps are problematic for an optimisation approach of allomorph distribution, since there could be an optimal output, and yet we find none.

Any scenario where a non-optimal affix appears is a further crux for a theory of optimised allomorphy. Bye (2008) gives several examples that are 'impossible to square with the standard assumption that allomorph distribution falls out from markedness considerations' (p.22) – that is, refute the notion that shapes of allomorphy will thrive to be optimal if possible.

One of his examples comes from Dyirbal, a Pama-Nyungan language spoken in Queensland, Australia (Dixon, 1972). In this language, the ergative marker has two allomorphs, *-gu* and *-ŋgu*. The latter attaches to the head foot (Dyirbal stress is trochaic), while the former appears elsewhere (cf. table 11).

(ya.ɽa-ŋ).*gu*	'man'
(ya.ma).(ni.-*gu*)	'rainbow'
(ba.la).(ga.ra).-*gu*	'they'

Table 11: Dyirbal ergative affixation (Bye, 2008, p.12)

A markedness-based approach would have the allomorph -*gu* as the winner in both cases, i.e. in the head-foot and elsewhere, as it is shorter and its infixation does not violate the universal constraint NoCoda. The only way the actual distribution can be extorted from an optimality-theoretic analysis is to grant -*ŋgu* a priority status, so that it does not have to compete with -*gu*. The latter is then only evaluated as a possible infix *if* the requirements for the infixation of -*ŋgu* are not satisfied.

Bye has several other examples of non-optimal affixation. He argues, following Paster (2005), that these should not be handled by phonological optimisation, but rather by a separate module of language-specific morphological subcategorisation. In such a model the templates for allomorph selection are stored separately, and are independent from the phonology.

The markedness-based approach to allomorphy only seems to work when one of the allomorphs is less marked in terms of complexity or phonotactics/prosody. It fails in cases of phonological discord, when a non-optimal allomorph appears in a certain position, refuting our expectations of naturalness. It also fails if there is no form at all for a particular morphological function, as with the Norwegian imperatives.

While an analysis which regards subcategorisation as language-specific and separate from phonological optimisation is able to deal with these cases, it has two drawbacks. Firstly, it puts extra weight on the grammar, not required in cases where allomorph distribution does seem to fall out from markedness considerations. Secondly, it does not explain the subcategorisation of the allomorphs, only states their nature. We could say that it is not the duty of a synchronic analysis to make sense of allomorph distributions, as these distributions result from a series of diachronic changes, and not all of these changes are acting in the direction of naturalness.

4.3 Naturalness and the Hungarian possessive

On the one hand, equating allomorph distributions with phonological optimisation has a large explanatory power not only when the positioning of one or more similar allomorphs is to be determined, but also in cases of suppletion, where the allomorphs are in no relation whatsoever with each other, but their placement reflects requirements of euphony (cf. section 4.1).

On the other hand, such an approach is less fruitful in cases where allomorph selection displays phonological discord, i.e. when the resulting forms are not the optimal ones. With this in mind, we turn to the Hungarian possessive, to see whether it is compatible with a markedness-based account.

The categorical behaviour of the 3SG-POSS does meet such an account. The categorical properties of its distributions are repeated below for the reader's convenience:

- The 3SG-POSS suffix (*-(j)A*) has four possible allomorphs: *-a, -e, -ja, -je.*
- V-final stems always have *-jA*.
- Sibilant/palatal C-final stems always have *-A*.

The effects of vowel harmony require no further considerations. The hiatus filler glide after vowel-final stems is a very frequent cross-linguistic pattern, especially since the hiatus filler is a glide. Hiatus filling is recognised as a repair strategy both to reach more natural syllable structure and to enhance perception (Clements & Keyser, 1983; Casali, 1996). Of course, the case of the possessive is slightly different: the glide is the only hiatus filler in Hungarian, but usually its appearance is phonological – it has a source in the neighbouring vowels. That is, it only occurs between two vowels if one of them is mid or closed (Siptár, 2008): compare *dió* [dijoː] 'walnut' and *faág* [fɒaːg] 'branch'. In the possessive, it always occurs between vowels, even if both are open: *almája* [ɒlmaːjɒ] 'apple-3SG-POSS'.

The avoidance of the glide after sibilant and palatal consonants is a strong pattern in Hungarian, as it was illustrated using the verbal imperative in section 2.2. It can also be blamed on the cross-linguistic tendency to avoid sequences of identical features, or even the phonetics. It is harder to parse a sequence of a palatal stop/nasal plus a glide perceptually, and it is also difficult to articulate a sequence of a palatal or a sibilant plus a glide (Hayes et al., 2004).

The variable patterns, however, are problematic. Recall that consonant-final stems do occur with the *-jA* form, albeit only in 9 per cent of the cases (as calculated on tokens), and that one can observe strong correlations between the form of the word and the presence or absence of the glide. The two most important are repeated below:

- Consonant-final back vowel stems prefer *-jA* in comparison with consonant-final front vowel stems (with a *-ja*:*-je* ratio of 87:13).
- Stems ending in a consonant cluster prefer *-jA* in comparison with stems ending in a single consonant (with a *-ja*:*-je* ratio of 67:33).

On the face of it, the Hungarian case is similar to the Dyirbal one in the sense that the less complex allomorph, *-A* should win out in all cases when the stem ends in a consonant, and the initial glide has no obligatory hiatus-filling function. This is illustrated by the partial OT analysis in table 12. The two constraints employed are *Hiatus, penalising vowel sequences, and *Struc, simply penalising structure. It expresses the notion that suffixes should be as short as possible, as allomorph selection (we think) is driven by euphony. We assume that there are further faithfulness constraints outranking these two in such a manner that hiatus filling (insertion) is allowed, but deletion of the stem is not. Vowel harmony is taken for granted.

The underlying representation has both possible allomorphs: this echoes Rebrus (2000), who posits two different morphemes for the possessive.

	/ke:p+{jɛ,ɛ}/	*Hiatus	*Struc
☞	ke:pɛ		*!
	ke:pjɛ		

	/kɒpu+{jɒ,ɒ}/	*Hiatus	*Struc
☞	kɒpujɒ		*
	kɒpuɒ	*!	

Table 12: Vowel-final stems taking *-jA*

This analysis works with vowel-final stems, such as *kapu* ‘gate’ and consonant-final stems which take the *-A* form, such as *kép* ‘picture’. We

run into difficulties with words like *nap* 'sun', which end in a consonant but consistently take the *-jA* suffix in all cases (table 13).

	/nɒp+{jɒ,ɒ}/	*Hiatus	*Struc
☞	nɒpɒ		
☹	nɒpjɒ		*!

Table 13: Consonant-final stems (should not be) taking *-jA*

Alternatively, we could use stochastic ranking (Boersma, 1998) to partially order *Hiatus and *Struc, and use a separate constraint to ban palatal/sibilant plus *-jA* sequences. This approach could get us the observed overall variation (9 per cent of C-final stems and none of the palatal/sibilant-final stems taking *-jA*), but the shortcomings are obvious.

The variation between *-jA* and *-A* does not necessarily occur at the word level. As we saw in section 3, most stems only take one suffix or the other, and only a subset of the stems show alternations between the two themselves (cf. section 3.3).

Furthermore, very similar stems categorically prefer one form or the other, as in the case of *oszlopa* 'column-3sg-poss' and *stopja* 'stop-3sg-poss'. Even if we regard this as accidental, we still have to tackle the observed sub-regularities, as listed above. It would not be easy to frame a phonological constraint that gives us the preference of *-jA* for back vowel stems or stems ending in consonant clusters. In this sense, then, it is impossible to put the weight of the distributions entirely on the shoulders of phonological optimisation.

It is a point of interest that *-(j)A* allomorphy also goes against the principles of Natural Morphology. Even though the *-jA* form is more marked than the *-A* one after consonant-final stems, this is the pattern favoured by the change. It spread to post-consonantal position, and now the C-final loanwords take the *-jA* form too. In this sense it is productive as well.

As Papp (1975) points out, having the *-ja* form after consonant-final words is better considering the alignment of word and syllable boundaries, allowing for a better discrimination of the stem, cf. *oszlo.pa* vs. *stop.ja*. This could explain the spread of this pattern to consonant-final environments, but not its preference after clusters or back vowel stems.

If phonological markedness is insufficient to describe the behaviour of the 3SG-POSS, we have to find further, external factors. The next section posits a diachronic explanation, heavily relying on the concept of analogy as the influence of similar forms in the speaker's lexicon.

5 *Diachrony and -(j)A* allomorphy

This section proposes an analogical explanation of *-(j)A* allomorphy. I will argue that the appearance of *-jA* after C-final stems is due to the effect of analogical relations which had their impact during language change. Analogy here is interpreted as a general influence of similar forms, irrespective of their function. Since this interpretation differs from the classical concept of analogy in linguistics, we will first briefly overview its properties.

5.1 Analogy in language change

The Neogrammarians regarded analogy as a counter-weight of sound change. Whereas sound change is affecting the trigger environments uniformly, blind to their possible morphological functions, and creates morphological irregularity, analogy acts irregularly, repairing paradigms altered by sound change, thus creating morphological regularity.

Analogy is, then, sensitive to the relationship between function and form, whereas sound change is not. A typical analogical levelling scenario will affect a particular paradigm, altering the forms in it to be more uniform. One example from McMahon (1994) is the levelling observed in the Modern English and German paradigms of the verb *choose*[4]. Both Old English and Old High German had alternations resulting from sound changes involving [s], [z], and [r]. These alternations are levelled out in the modern languages, giving a uniform [z] in English and [r] in German (cf. table 14).

The *choose*-example is instructive: analogical levelling takes place within one paradigm, and it resolves the disconcertion left by sound change. Analogical levelling outside paradigms should only occur sporadically, in cases of Contamination. One of these is the effect of *brother* on *father* in English, the latter having a medial [ð] instead of a [d] due to the influence of the former.

A similar example is the Hungarian verb form *könyörgöm* 'plead-1SG-INDEF'. This indefinite form is identical to the definite. This is a property

[4] *Küren* is dated in German.

Function	Old English	Modern English
present	cēo[z]an	choo[z]e
past sg.	cēo[s]	cho[z]e
past pl.	cu[r]on	cho[z]e
past part.	geco[r]en	cho[z]en
Function	Old High German	Modern German
present	kiu[s]an	kü[r]en
past sg.	ko[s]	ko[r]
past pl.	ku[r]un	ko[r]en
past part.	giko[r]an	geko[r]en

Table 14: Analogical levelling in Germanic

of some verbs with a 3SG-INDEF ending *-ik* in Hungarian, which take their 1SG-DEF as their 1SG-INDEF, e.g. *eszik* 'eat-3SG-INDEF' – *eszem* 'eat-1SG-INDEF', *alszik* 'sleep-3SG-INDEF' *alszom* 'sleep-1SG-INDEF'. (Compare with *megy* 'go-3SG-INDEF' – *megyek* 'go-1SG-INDEF'.) The verb *könyörög* has no *-ik* ending in 3SG-INDEF, nonetheless, it has an indefinite ending in *-m*, most likely due to the influence of the *-ik* verbs[5].

In short, analogy in language change is usually regarded as a force going against sound change, and, in the case of levelling, acting within paradigms. While examples of 'containment' across paradigms are to be found, they are considered accidental.

Another view of analogy is one where it simply means that similar forms exert an influence on each other, irrespective of their functions, and by virtue of it, they become more similar. The preference for similarity is grounded in concepts of economy of perception and economy of storage. To put it simply, fewer differences are easier to remember. Preference for similarity is counter-weighed by the avoidance of homonymy, so that forms marking different functions will not become similar, unless if they are distinguishable in some other way[6].

[5]Hungarian speakers are aware of this, and keep correcting each other when they say *könyörgöm*, regarding *könyörgök* as the 'correct' form. It is also true that, for most speakers, the *-m* form only functions as a discourse particle, not as a verb.

[6]A similar concept in the literature on categorisation is the *perceptual magnet effect* (**?**):

This might sound vague, but can be exemplified by the Hungarian infinitive (Kálmán & Rebrus 2010; Siptár 2009). The infinitive in Hungarian can be marked for person and number. A partial paradigm is shown in table 15.

Inf	alud*ni*	emel*ni*
1SG-INF	alud*nom*	emel*nem*
2SG-INF	alud*nod*	emel*ned*
3SG-INF	alud*nia*	emel*nie*
Gloss	to sleep	to lift

Table 15: The Hungarian infinitive

As Hungarian usually displays clear agglutinative word formation with distinguishable affixes, we would expect such a pattern for the INF throughout the paradigm: the infinitive ending *-ni* plus a person/number marker (1SG *-m*, 2SG *-d*, 3SG *-a/-e*, etc.). The intriguing thing is that instead of the expected **aludnim*, **aludnid*, we find *aludnom* and *aludnod* for the 1/2SG-INF.

Rebrus and Kálmán claim that this is an analogical levelling effect. To take the example of *aludnom* 1SG-INDEF, there is an overwhelming majority of nominal possessive forms ending in *-om/-öm, -od/-öd* (six thousand versus three hundred in the Hungarian Webcorpus, and the ratio is the same with nominal forms in general).

Since, as we will shortly see, the nominal possessive paradigm is very similar to the verbal one, a small difference within possessive nouns (*-nom* instead of *-nim*), reinforced by numbers, can have an effect on the formation of the infinitive.

The difference between this type of analogical levelling and levelling in its classical sense is that this levelling takes place across paradigms (from the nominal possessive to the verbal infinitive) and is oblivious to function: the only thing the two paradigms share is formal similarity.

The nominal paradigm can intrude on the verbal one as there is no chance to confuse a verbal infinitive with a nominal possessive, given the scarcity of similar stems, the sharp semantic distinctions, and the different syntactic

instances assigned to a category will be perceived as more similar to the category centre, creating local maxima in the long run.

distributions. The infinitive forms can become more similar to the nominal ones and therefore more economical to store without the cost of homonymy.

5.2 Analogy in the possessive

The concept of analogy as discussed above can be applied to the Hungarian possessive. I will argue that we are witnessing a similar analogical levelling scenario as in the case of the verbal infinitive, except that we have more factors to reckon with.

The first of these factors is the above mentioned similarity between the nominal possessive and the verbal paradigms. The singular definite and the plural 1/2 indefinite parts of the verbal paradigm are quite close to the corresponding nominal forms[7] (Rácz & Rebrus, 2010). This development is hardly surprising if we assume the view of analogy suggested above: the two paradigms are difficult to confuse, so they can become identical to spare space. Table 16 illustrates the similarity.

	Front		Back	
	Noun.Poss	Verb.Def	Noun.Poss	Verb.Def
	kert	ért	part	tart
1sg	kert*em*	ért*em*	part*om*	tart*om*
2sg	kert*ed*	ért*ed*	part*od*	tart*od*
3sg	kert*je*	ért***i***	part*ja*	tart***ja***
	Noun.Poss	Verb.Indef	Noun.Poss	Verb.Indef
1pl	kert*ünk*	ért*ünk*	part*unk*	tart*unk*
2pl	kert*etek*	ért*etek*	part*otok*	tart*otok*
3pl	kert*jük*	ért*jük*	part*juk*	tart*juk*
Gloss	'garden'	'understand'	'riverbank'	'hold'

Table 16: The nominal possessive and the verbal paradigms

The similarity is apparently strong enough to allow the numerically superior nominal possessive to exert its influence on the verbal infinitive. This

[7]The last row of this table is apparently misleading. It is true that the 3pl of the nominal possessive is similar to a verbal form, but it is not the corresponding 3pl-indef, but the 1pl-def! (*kertjük* 'their garden' vs. *értjük* 'we understand it')

might be possible the other way around as well. Recall from section 3 that one strong pattern in *-(j)A* variation was a preference for back V stems. If we look at table 16, we can see that one difference between the two paradigms is that the verbal *3sg* has a *-ja* ending in the back paradigm but an *-i* ending in the front one.

What follows is that if there is any effect on the distribution of *-(j)A* from the similar verbal paradigm, it will only show with back stems, as the front stem ending is completely different. This is precisely what we find: more back vowel noun stems have the *-ja* form, since there is a corresponding verbal *-ja*. This factor is absent in the case of the front vowel forms, as the ending *-i* is not an option in 3SG-POSS.

The direction of the levelling is odd in the sense that it is not reinforced by numbers. In the case of the infinitive, the nominal paradigm could meddle in the verbal one because it was much more populous. Obviously, this cannot work the other way around as well.

Then again, there are significant differences between the two cases. The verbal ending is categorically present, whereas the nominal ending is variable. We could say that the similar verbal forms have an effect because the nominal forms are underdetermined in 3SG-POSS. Indeed, if the analogical source was in a strong majority, we would expect a complete levelling of the nominal possessive, so that we found *-ja* in all forms. That is to say, the effect of the verbal forms is present, but it is not overwhelming. Since the verbal endings are fixed, there is no possibility for the nominal paradigm to influence the verbal one.

Another important factor is perceptibility. The ending *-je* is much harder to perceive than *-ja*, since its segments are more similar (Marilyn Vihman, p.c.). No wonder that the verbal 3SG-DEF has a different ending for the front vowel forms. It is important to note though that this is also just a tendency. We do find C-final nominal forms ending in *-je*, and, as a matter of fact, the verbal subjunctive ending for the 3SG is also *-je* (as in *emelje* 'lift-3SG-SUBJ' – the back vowel ending is *-ja*. The token frequencies of these, however, is quite small.

The correlation between stems ending in a consonant cluster and the *-jA* form is explored in detail by Papp (1975), who claims that the preference for the glide-initial ending derives from a tendency to avoid homonymy, as a large number of nominal types end in *CCa/CCe*.

These are the main factors shaping the variation of the nominal 3SG-POSS. What we essentially have is a situation where the form corresponding to

a certain function (the 3SG-POSS) is underdetermined and unstable. Since there is no evident choice for a 3SG-POSS form, sub-regularities and hidden patterns emerge. Perceptibility and the knowledge of similar forms can have a direct effect on unstable variation, the latter both in the sense of homonymy avoidance and preference of similarity.

5.3 A note on representations

As we saw, the variation of the 3SG-POSS after C-final stems is impossible to squeeze out of phonological optimisation. Therefore, it has to be assigned in the lexicon as a subcategorisation frame, in the sense of Bye (2008); Paster (2005). This still leaves us with problems, though. Very similar words can go for one allomorph or the other, and we find patterns of correlation with vowel quality, consonant clusters, final consonant quality, etc., which would be very difficult to compress into a subcategorisation frame.

A minimal memory model of lexical storage would want to derive the allomorph distributions from matrices of underlying forms and a few relatively simple templates of their concatenation. This is possible for the categorical distributions of the possessive. We can imagine lexical representations for the suffixes such as those in table 17.

-ɒ	**-ɛ**	**-jɒ**	**-jɛ**
3SG-POSS	3SG-POSS	3SG-POSS	3SG-POSS
back V stem	front V stem	back V stem	front V stem
C-final stem	C-final stem	V-final stem	V-final stem

Table 17: Minimal memory model representations of 3SG-POSS allomorphs

Such a representation would still suffice if the words displayed general variation, but none at the word level. Unfortunately, we find word-level variation. The only way to integrate all this information into a minimal memory storage would be to supply individual words with different matrices, which are also incorporating information like amount or quality of final consonants.

An alternative to this is the use of an exemplar-based representation (Nosofsky, 1988; Johnson, 2005; Bybee, 2006). The exemplar-based lexical representation of a word form consists of its previously recorded utterances, containing phonetic detail. The behaviour of word forms is basically determined by analogical extension: if a word form A shares a number of properties

with word form B, and word form B has additional properties, then word form A will tend to acquire these properties as well. In exemplar models, this extension is usually modelled as a categorisation task: if the word A shares sufficient categories with category B, it is categorised as a member of this category, and follows the behaviour of the forms therein (taking a particular past-tense form, for instance).

The direction of the extension is usually determined by frequency, more frequent types having an effect on less frequent types. This was the example we saw with the infinitive: nominal possessives are generally similar to verbal infinitives, and they are the majority, so verbal forms will become even more similar to nominal ones in the long run. In our case, the fact that the nominal 3SG-POSS is unstable, whereas the similar verbal 3SG-DEF is not, changes the direction of the extension.

In an exemplar-based representation, a word form is the set of its exemplars, previously recorded utterances. Inflected forms do not result from concatenation, they are simply stored as well, and the link with the non-inflected form (as in *cipő–cipője* 'shoe'–'shoe-3SG-POSS') is established through semantic and phonetic similarity.

Such a model can store the word forms individually, allowing for variation between the words. The detailed representations allow for the extraction of any pattern of similarity, including the preference of *-jA* for back vowels or CC-final words. What follows is that an exemplar-based representation is more suited to describe *-(j)A* allomorphy than a minimal model one.

The analogical relations posited above go well with such a memory model. No wonder than that most models of analogy assume a rich memory representation, cf. Skousen 2002; Sóskuthy 2009). Recorded frequency of utterance allows us to establish the direction of the analogical extension, while phonetic detail permits the comparison of word forms in a trivial manner. Phonological and morphological patterns can emerge from the lexicon without the need for an explicit set of rules or constraints (though it is debated whether they do so, cf. Becker et al. 2007; Hayes et al. 2009).

5.4 Analogy as a diachronic explanation

Above I proposed a possible explanation of the distributional patterns of the possessive, building on the concept of analogy as the influence of similar forms and an exemplar-theoretic representation. To sum it up again, the 3SG-POSS is unstable, that is why it shows variation. Since there is no evident choice of

form for this function, various sub-regularities can emerge and influence it.

Needless to say that this is not an online process. Speakers need not compare paradigms in their heads before they utter a possessive nominal form. The factors listed above exerted their influence during language change. Recall from section 2.2 that the glide-initial allomorph was first restricted to post-vocal environments, and slowly spread to post-consonantal ones. It is hard to say what triggered this change, and possibly all the above factors played a part in it, though we can risk to say that the main initiator was the need of phonetic distinctiveness. The change, as soon as it started, was fuelled by analogy on the basis of the similar verbal paradigm, as well as the factors of homonymy avoidance.

What we end up with is a synchronic pattern with a diachronic explanation, which is in line with a rising trend in phonology and morphology (Blevins, 2004; Pierrehumbert, 2001). The diachronic account renders a synchronic one redundant in the sense that if we can explain the development of a pattern we do not need to explain its shape once again. Even putting that aside, there is no available, streamlined alternative to this approach in describing the Hungarian possessive, as *-(j)A* variation cannot be solely based on phonological optimisation or morphological naturalness.

6 Simulating possessive distributions

The previous section, in line with the rest of the paper, argued that the observable variation of the 3SG-POSS after consonant-final stems reflects its under-determinedness. Since there are no categorical patterns determining whether most C-final stems should take *-jA* or *-A*, stark tendencies are able to emerge, the strongest of them being the effect of the similar verbal paradigm.

In this section I discuss the results of a simulation which regarded possessive allomorphy as a categorisation task. The basic reasoning is as follows: if allomorph selection is categorical, it presents in strong patterns that a categorisation algorithm can easily extract, so that these patterns can be re-created, and the result of the categorisation task will almost strictly match its input. If, however, allomorph selection shows variation, this variation takes its toll on pattern learnability, and the result of the categorisation will show more errors.

The modelling of possessive variation is performed using the Tilburg Memory-based Learner (TiMBL) (Daelemans et al., 2007). TiMBL is an

algorithm categorising items into classes based on their characteristics. It starts with the assumption that the behaviour of a form can be derived from the behaviour of similar forms. That is to say, if a certain noun takes the *-jA* suffix in the possessive, similar nouns will do so too. TiMBL's input is typically a pair of corpora: one input corpus and one test corpus. The input corpus consists of word forms defined by features. The last one of these features is the one that has to be predicted – in this case, whether a word takes *-jA* or *-A* in 3SG-POSS. The test corpus also consists of forms, with the difference that the feature the algorithm should predict is missing.

First, the algorithm calculates the information gain of the input features. The information gain of a particular feature depends on the ability of the feature to predict the feature we want to get. For example, if a noun stem ends in a vowel, it will certainly take *-jA*, so in this sense the last segment of the word, as a feature, is a very good predictor. In comparison, the first segment of the word is pretty much irrelevant from the point of view of allomorph selection. If we take the segments of the word form as features, the last segment will be a much better predictor of the allomorphy than the first one.

Second, the algorithm establishes the distance between the forms in the input corpus and the test corpus. In doing so, it takes into account the amount of similar features two forms share, as well as the information ratio of these features. Finally, it assigns the missing feature to the forms of the test corpus, based on their distance from the forms of the input corpus. So, if a test form is, in general, more similar to input forms that select *-jA*, it will also select *-jA*. The success of the categorisation task depends on the strength of the correlation between the given features and the feature we want to predict. If words of a particular form consequently select one allomorph, the algorithm will successfully categorise a novel, similar word as taking this allomorph. If allomorph selection has a large number of sub-regularities and exceptions, categorisation will be less successful. TiMBL has other, different settings of operation, but the above ones were used in the present task.

In the previous section we assumed that 3SG-POSS variation results from the fact that speakers are unable to consequently select one allomorph or the other, lacking categorical patterns which would 'instruct' them to do so. If that is so, we have two predictions on a categorisation task on possessive allomorphy. First, there will be errors in the results of the categorisation (as the input is under-determined). Second, the tendencies observed in the Webcorpus will be, to an extent, re-created. We will see that this simulation

confirms both suspicions.

I performed two tests of categorisation. The first included both definite verbs and possessive nouns in the input, and was tested on nouns only. The second one had only nouns in the input and was confined to consonant-final stems both in the input and the output.

6.1 Test 1

6.1.1 Procedure

Two thousand nominal and verbal forms were extracted from the Webcorpus to create the input. The sampling was random, so the ratios of various forms (as well as nouns and verbs) should reflect the general ratios of the corpus.

The extracted forms are types. Working with TiMBL requires using types, because it calculates the distance of the target form from nearest neighbours – if we use tokens, the nearest neighbours will be identical to the target form, undermining the success of categorisation. In section 3, I emphasise the importance of token frequency versus type frequency in the patterning of possessive variation. One of the arguments for concentrating solely on the former is that there are much more back vowel types in the 3SC-POSS set, but this asymmetry disappears if we look at tokens.

In order to avoid the skewing effect of type frequency, the input was sampled from words with a token frequency above one hundred in the corpus. This is an arbitrary threshold, but it approximates the point in the frequency curve where the ratio of types reflects the ratio of tokens. Since the asymmetries like the one above are caused by types with low token frequency, the token frequency cut-off filters them out.

Distance was calculated based on five features: the part-of-speech tag of the form (N or V), and the last four letters. (Hungarian orthography is relatively consistent, so using phonetic transcription is not necessary.) The feature to predict was whether the ending (*-(j)A*) starts with a glide. The forms were extracted from the corpus along with their corresponding 3SG-POSS or 3SG-DEF pairs, and the presence or absence of the glide was based on the latter. The last four letters were used so that both preceding vowel quality and the presence or absence of consonant clusters could be a defining factor in categorisation. The input is illustrated by table 18.

The test set was similar to the input set, except that it contained only four hundred forms. The categorisation task was to determine whether a

Features	Word	Gloss
N,r,r,á,s,n	forrás	‘source’
V,a,s,o,l,j	javasol	‘to suggest’
N,h,a,n,g,j	hang	‘sound’
V,s,g,á,l,j	vizsgál	‘to inspect’
N,á,r,g,y,n	tárgy	‘object’

Table 18: TiMBL inputs with features 1-5 plus feature 6, the feature to be predicted

particular form would take *-jA* – in the possessive if it is a noun and in the definite if it is a verb.

6.1.2 Results

The algorithm was able to categorise 95.75 per cent of the test forms (383 from 400) correctly. This suggests that the nouns and the verbs together present a robust, learnable pattern for the endings. The weighting of the features for the first test is shown in table 19.

Gain ratio is a weighted value of the information ratio: features with many values have a better information ratio, but can actually be worse predictors. Gain ratio is weighed with the number of values a feature can have. This can be seen with feature 1, word class: it has only two values, but it is actually a very good predictor of the ending.

Features	Values	Information gain	Gain ratio
1	2	0.16854417	0.26471666
2	31	0.069968913	0.015253455
3	32	0.18546907	0.043594327
4	29	0.35151535	0.092925131
5	29	0.53380254	0.14704314

Table 19: TiMBL feature values and weights for test 1

As we can clearly see in table 19, the best predictor of the ending is the first feature, word class, with a gain ratio of 0.26. It is followed by the fifth

feature, the last letter of the word, with a gain ratio of 0.15. This is hardly surprising, as both verbs (F1) and words ending in vowels (F5) categorically take a glide-initial ending, while sibilant/palatal final nouns categorically avoid it. The other features are much less important in allomorph selection.

6.2 Test 2

6.2.1 Procedure

The input for the second test consists of 1200 randomly sampled nouns, all consonant-final. The features are identical to those of the first test (even feature 1, to preserve a uniform layout for the data sets.) The test consists of 500 consonant-final nouns. The size of the test set was expanded to allow for a more detailed view of the possible categorisation errors.

6.2.2 Results

The algorithm categorised 77.20 per cent of the test forms (386 from 500) correctly. This is a twenty per cent decrease in comparison with the first test, suggesting that the consonant-final nominal patterns are not perfectly stable. This is in line with our earlier observations on the variation of this set in the Webcorpus.

Features	Values	Information gain	Gain ratio
1	1	0.0000000	0.0000000
2	31	0.057103311	0.012781310
3	29	0.10928759	0.029886376
4	19	0.15878056	0.057465263
5	14	0.19481348	0.074839964

Table 20: TiMBL feature values and weights for test 2

As table 20 shows, feature 5, the last letter of the word is a worse predictor in test 2, where vowel-final forms are missing. The weight of defining the ending pattern is more evenly shared between features 4 and 5. In general, however, word endings augur badly to the success of categorisation in the absence of verbs and vowel-final nouns – the gain ratios are smaller. This again illustrates the under-determinedness of the 3SG-POSS with C-final stems.

It is instructive to look at the error patterns in the test set. The three subsets in which error rates soar are words ending in a dental stop, words ending in a consonant cluster, and back-vowel forms. Recall from section 3 that these are the loci of phonologically motivated sub-regularities, namely, the preference of the glide-initial possessive suffix.

The two aspects in which these three subsets are interesting are error rates and categorisation patterns. If forms belonging to a certain subset are miscategorised to a large extent, we can say that these errors reflect variability – a form belongs to one set, but based on its shape, it could might as well belong to the other. A separate question is that of categorisation: we saw, for instance, that back vowel stems prefer *-jA* more than front vowel stems. If this is a learnable pattern, we except it to recur in the results of the categorisation task.

All three subsets show larger error rates and preference for -JA. A chi-square test of significance shows $p<0.001$ for the distributions in all three cases.

	n→ j	j→ j	j→ n	n→ n	Total
Dental ending	82	31	28	114	255
No dental ending	0	1	2	242	245
Total	82	32	30	356	500

Table 21: Results with stems ending in a dental stop

Table 21 shows the categorisation results projected to word endings, i.e./ whether a stem ended in a dental stop or not. The four columns indicate categorisation results. For example, 'n→ n' means that a form not selecting a glide-initial allomorph was correctly categorised to do so, while 'n→ j' means that a form not selecting a glide-initial allomorph was erroneously categorised as selecting one, etc. So, in the first case, *hát* 'back' was categorised as having the 3SG-POSS *háza*, in the second case it was categorised as going for **hátja*.

Two patterns emerge in this data set. First, stems ending in a dental stop overwhelmingly prefer *-jA* in the possessive. Almost all of the *-jA* forms in the results end in a dental stop. Second, the error rates are very high in the subset. A majority of the stems ending in dental stops are miscategorised (see column 'nj' in table 21). This partly shows (again) the preference of these forms for *-jA*. What it also shows is that it is difficult to predict the

behaviour of these forms: TiMBL simply miscategorises a lot of them, as the input patterns are not strong enough to determine class membership (*-jA* or *-A*) completely.

	n→ j	j→ j	j→ n	n→ n	Total
VCC-final stem	9	25	17	5	56
VC-final stem	73	7	13	351	444
Total	82	32	30	356	500

Table 22: Results with stems ending in a consonant cluster

Table 22 shows the results projected to stem endings, now in the sense whether the stem of the target form ends in a single consonant or a cluster. The main observations of the previous subset still stand here. On the one hand, the majority of CC-final stems take the *-jA* form. On the other hand, the majority of these stems is also miscategorised.

	n→ j	j→ j	j→ n	n→ n	Total
Back vowel	65	25	19	148	257
Front vowel	17	7	11	208	243
Total	82	32	30	356	500

Table 23: Results with back vowel stems

The last subset under scrutiny is that of back-vowel stems. Table 23 shows the distributions of the categorisation task projected to stem vowel quality. The patterns of larger error rates and preference for *-jA* repeat themselves.

The possible conclusion of the three examined datasets is that the patterns found in the Webcorpus have been re-created during the categorisation task, and that the selection of the glide-initial allomorph is highly variable. This is reflected in the error rates it displays in all three cases.

In sum, the simulation aptly illustrated that 3SG-POSS allomorph selection is variable, as in a large number of cases there is no trivial candidate for the 3SG-POSS of a particular form. It also showed that the observations drawn on Webcorpus data present in patterns learnable from a sampled dataset.

The validity of the observations on the categorisation results is grounded by the fact that it managed to reproduce other patterns familiar from the Webcorpus data. We could say that the error rates display another side of the variation not directly observable there. In this sense, then, the simulation is not only useful in reproducing the Webcorpus data, but it also sheds light on the nature of variation in ways unseen in a corpus study.

At this point it is wise to trace our steps back to the concept of exemplar-based lexical storage. TiMBL uses a prototype-based model of storage, where every word form has its distinct entry. Suffixation patterns are extracted from a dataset of pairs of suffixed and non-suffixed forms – it is not present as an abstract rule or schema. This way, TiMBL was able to reproduce most of the variation we observed in section 3. Furthermore, it was precisely this way of storing information that allowed TiMBL to fail in certain cases, which would hint at similar causes of variation in natural speech: if similar forms in the lexicon do not show strong enough suffixation patterns, allomorph selection will be more variable.

7 Conclusions

This paper gave a description of 3SG-POSS in Hungarian. I showed that its distribution is a case of phonological discord, that is, allomorph selection does not meet considerations of phonological markedness or naturalness. This is problematic for analyses that try to ground allomorph selection solely in phonological optimisation.

Furthermore, the 3SG-POSS displays intricate patterns of variation after C-final stems, and integrating these patterns into a synchronic grammar of allomorph selection borders on the impossible. The alternative offered in this paper is analogical influence exerted by similar forms in the lexicon, affecting the possessive in the course of language change.

By providing possessive variation with feasible explanations drawn from the diachronic domain, this paper supplies evidence for the validity of diachronic explanations. It also supports the idea of rich memory models of lexical storage, as the established analogical relations rely on token frequency and detailed similarity of form, information that a minimal memory model cannot offer.

Using analogy across paradigms as a culprit for the variation displayed by one particular morphological function expands the number of possible factors

in allomorph selection, and hopefully widens our understanding of change in morphology.

References

Anderson, S. R. (1972). On nasalisation in Sundanese. *Linguistic Inquiry* 3, pp. 253–268.

Anttila, A. (2002). Variation and opacity. *Natural Language and Linguistic Theory* 20, pp. 1–42.

Anttila, A. (2007). Variation and optionality. *The Cambridge Handbook of Phonology*, CUP.

Becker, M., N. Ketrez & A. Nevins (2007). The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish devoicing neutralization. Ms. UMass Amherst.

Blevins, J. (2004). *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge: Cambridge University Press.

Boersma, P. (1998). *Functional Phonology*. Ph.D. thesis, University of Amsterdam.

Bybee, J. (2001). *Phonology and Language Use*. Cambridge: Cambridge University Press.

Bybee, J. (2006). *Frequency of Use and the Organization of Language*. OUP.

Bye, P. (2008). Allomorphy - selection, not optimization. Blaho, S., M. Krämer & P. Bye (eds.), *Freedom of Analysis*, Berlin & New York: Mouton de Gruyter, p. 63=91.

Carstairs-McCarthy, A. (1988). Some implications of phonologically conditioned suppletion. *Yearbook of Morphology* pp. 67–94.

Casali, R. F. (1996). *Resolving Hiatus*. Ph.D. thesis, UCLA.

Clements, G. N. & S. J. Keyser (1983). *CV Phonology – a Generative Theory of the Syllable*. MIT Press.

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

Dixon, R. M. W. (1972). *The Dyirbal Language of North Queensland*. CUP.

Dressler, W. U. (1999). What is natural in natural morphology (NM)? *Prague Linguistic Circle Papers* 3.

Dressler, W. U., W. Mayertaler, O. Panagl & W. U. Wurzel (eds.) (1987). *Leitmotifs in Natural Morphology*. Amsterdam: John Benjamins.

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

Hayes, B. & Z. Cziráky Londe (2006). Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology* 23, pp. 59–104.

Hayes, B., R. Kirchner & D. Steriade (2004). *Phonetically Based Phonology*. Cambridge: Cambridge University Press.

Hayes, B., K. Zuraw, P. Siptár & Z. Londe (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85, pp. 822–863.

Johnson, K. (2005). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *UC Berkeley Phonology Lab Annual Report*.

Kálmán, L. & P. Rebrus (2010). Valóban megmagyarázhatatlanok a magyar infinitívusz toldalékai? [Are the Hungarian infinitive suffixes really inexplicable?]. *A mai magyar nyelv leírásának újabb módszerei*, VII. SZTE, Szeged.

Zsuzsa Kertész & L. Kálmán (2008). Russian palatalisation. *Papers from the Mókus Conference*, 83-92, Budapest: Tinta.

Kiefer, F. (1985). Natural morphology. *Acta Linguistica Hungarica* 35, pp. 85–105.

Kiefer, F. (ed.) (2000). *Strukturális magyar nyelvtan 3. Morfológia*. Budapest: Akadémiai Kiadó.

Labov, W. (2006). A sociolinguistic perspective on sociophonetic research. *Journal of Phonetics* 34, pp. 500–515.

McCarthy, J. J. (1986). OCP effects: Gemination and antigemination. *Linguistic Inquiry* 17, pp. 207–263.

McMahon, A. M. S. (1994). *Understanding Language Change*. CUP.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14(1), pp. 54–65.

Papp, F. (1975). *A Magyar Főnév Paradigmatikus Rendszere [The Paradigmatic System of the Hungarian Noun]*. Akadémiai Kiadó, Budapest.

Paster, M. (2005). Subcategorization vs. output optimization in syllable-counting allomorphy. Alderete, J., C. hye Han & A. Kochetov (eds.), *Proceedings of the Twenty-Fourth West Coast Conference on Formal Linguistics*, Somerville, MA: Cascadilla Proceedings Project, pp. 326–333.

Pierrehumbert, J. (2001). Why phonological constraints are so coarse-grained. *Language and Cognitive Processes* 16 (5/6), pp. 691–8.

Prince, A. & P. Smolensky (1993). Optimality Theory: Constraint interaction in generative grammar. Tech. rep., Rutgers University.

Rácz, P. & P. Rebrus (2010). Complexity and distinctiveness in the possessive allomorphy of Hungarian. *Talk given at OCP7 Nice*.

Rebrus, P. (2000). Morfofonológiai jelenségek és a Lexikon [Morphophonological processes and the Lexicon]. Kiefer, F. (ed.), *Strukturális Magyar Nyelvtan 3. Morfológia [Structural Hungarian Grammar 3. Morphology]*, Budapest: Akadémiai Kiadó, pp. 763–947.

Rebrus, P. & M. Törkenczy (2008). Morfofonológia és a lexikon [morphophonology and the lexicon]. *Strukturalis Magyar Nyelvtan IV. A Lexikon*, Budapest: Akadémiai Kiadó.

Rebrus, P. & M. Törkenczy (2010). Covert and overt defectiveness in paradigms. Rice, C. & S. Blaho (eds.), *Modeling Ungrammaticality in Optimality Theory*, Equinox Publishing.

Rice, C. (2005). Optimal gaps in optimal paradigms. *Catalan Journal of Linguistics* special issue on Morphology in Phonology.

Rubach, J. & G. E. Booij (2001). Allomorphy in optimality theory. *Language* 77, pp. 26–60.

Rumelhart, D. E. & J. L. McClelland (1987). *Parallel Distributed Processing - Vol. 1 Foundations*. MIT Press.

Siptár, P. (2008). Hiatus resolution in Hungarian: An optimality theoretic account. nón, C. P. & S. Szentgyrgyi (eds.), *Approaches to Hungarian 10: Papers from the Veszprém Conference*, Budapest: Akadémiai Kiadó, pp. 187–208.

Siptár, P. (2009). Morphology or phonology? the case of hungarian -ni. den Dikken, M. & R. M. Vago (eds.), *Approaches to Hungarian 11: Papers from the 2007 New York Conference*, Amsterdam: John Benjamins, pp. 197–215.

Skousen, R. (2002). *Analogical Modeling: An Exemplar-Based Approach to Language*. Amsterdam: John Benjamins.

Sóskuthy, M. (2009). *Analogy on the level of phonology – The emergence of intrusive-r in English*. Master's thesis, Eötvös Loránd University.