Acta Linguistica Academica Vol. **66** (2019) 4, 601–620
DOI: 10.1556/2062.2019.66.4.5

Frequency and prototypicality determine variation in the Hungarian verbal 1SG.INDEF

Péter Rácz
Central European University
raczp@ceu.edu

Abstract: I provide a synchronic account of the variation between the **marked** and **unmarked** forms of the 1SG.INDEF of Hungarian (-ik) verbs; verbs that end in (-ik) in the 3SG.INDEF. I use a generalised mixed-effects regression analysis to explore how these forms vary in an extensive sample of the language, the Hungarian Webcorpus. I find that verbs' preference for the marked/unmarked form is determined by their lemma frequency and their prototypicality as members of the (-ik) class. These results are consistent with a morphological levelling account of variation in Hungarian verbal morphology, in which verbs migrate away from the minority (-ik) class and into the majority regular class. This suggests a picture of variation in Hungarian verbs that is shaped by lexical organisation, morphophonology, and social dynamics.

Keywords: corpus linguistics; morphology; language variation and change; Hungarian

1. Variation in the 1SG.INDEF of Hungarian (-ik) verbs

Classic Literary Hungarian distinguishes two verb classes, regular verbs and (-ik) verbs. The distinction was, originally, at least partially **semantic**, between **active** and **medial** verbs. The semantic grounding survives in a handful of forms, such as [tør] 'break-TR.3SG.INDEF' and [tør-ik] 'break-INTR.3SG.INDEF'. At the same time, semantic category boundaries have been porous and it is unclear whether the system, partly re-introduced during the language reform period of the 19th century, has ever existed in its entirety in spoken language (Kiss & Pusztai 2003).

In any case, the Classic Literary Hungarian distinction between regular and (-ik) verbs is manifest at several points in the inflectional paradigm, as shown by Table 1. (Transcriptions do not mark non-lexical phonological processes to allow for better comparison across variants. Glosses follow the Leipzig Glossing Rules, see Bickel et al. 2008.)

Table 1: Differences between the verb classes in Classic Literary Hungarian shown by regular verb [feːʃyl] 'comb-3SG.INDEF' and (-ik) verb [feːʃylkødik] 'comb-3SG.REFL.INDEF'

	Regular verb			(-ik) verb		
	1SG.INDEF	2SG.INDEF	3SG.INDEF	1SG.INDEF	2SG.INDEF	3SG.INDEF
IND	feːʃyløk	feːʃyls	feːʃyl	feːʃylkødøm	feːʃylkødøl	feːʃylkødik
IMP	feːʃyljɛk	feːʃylj	feːʃyljøn	feːʃylkødjɛm	feːʃylkødjel	feːʃylkødjek
COND	feːʃylnek	feːʃylnel	feːʃylnɛ	feːʃylkødnem	feːʃylkødnel	feːʃylkødnek

Educated Colloquial Hungarian (Siptár & Törkenczy 2000) only maintains two of these paradigm distinctions, indicated in grey in the table above. These are the 3SG.INDEF indicative and the 1SG.INDEF indicative. The (-ik) class has an (-ik) suffix instead of (-∅) in the former and (-m) instead of (-k) in the latter. Otherwise, the regular and the (-ik) class are identical. Since (-m) is also the 1SG.DEF ending, (-ik) stems neutralise the contrast in **definiteness** in this position.

This is shown in (1)–(2): [tud] 'know.3SG.TR' is a regular verb. It shows a contrast between the indefinite (1a) and the definite (1b). On the other hand, [ɛsik] 'eat-3SG.INTR' is an (-ik) verb, which uses (-m) in both cases (2a–b).

(1) a. [tudok ɛɟ pompaːʃ vitsːɛt] 'I know a fantastic joke.'
b. [tudom ɒmit tudok] 'I know what I know.'

(2) a. [ɛsɛm ɛɟ kiʃ kɛɲɛrɛt] 'I eat a little bread.'
b. [ɛsɛm ɒ kɛɲɛrɛt] 'I eat the bread.

However, the first person singular indefinite (1SG.INDEF) of the (-ik) verbs shows variance between the neutralising (-m) variant and the "regular" (-k) variant (see (3)). I will refer to the (-m) variant as **marked** and to the (-k) variant as **unmarked**. This is because the (-m) variant **marks** the (-ik) stem in the 1SG.INDEF and because it **neutralises** the contrast between the 1SG.INDEF and the 1SG.DEF.

(3) a. [itː lɒkom pɛʃtɛn] 'I live here in Budapest.'
b. [itː lɒkok pɛʃtɛn] 'I live here in Budapest.'

Apart from a set of hypercorrect, lexicalised forms (such as [køɲørgøm] 'I implore you', cf. [køɲørøg] 'implore.INTR.3SG.INDEF'), the variation

between marked (-m) and unmarked (-k) is restricted to the 1SG.INDEF of the (-ik) verbs only.

This variation is a socially salient linguistic marker (Kontra & Váradi 1997). The marked form is seen as the "correct"/prestige variant and shows both social stratification and style shifting.

When we compare Classic Literary Hungarian with Educated Colloquial Hungarian, we witness the retreat of the (-ik) conjugation paradigm, which is effectively restricted to 3SG.INDEF and 1SG.INDEF in the latter. This, coupled with the social salience of variance in the 1SG, could indicate that the (-ik) paradigm is gradually disappearing (even if the (-ik) ending of the 3SG.INDEF remains stable).

This is reminiscent of the process of **morphological levelling**, in which verbs migrate from the minority category (the (-ik) class) to the majority category (the regular -∅ class) (Bybee 1985). Levelling does not necessarily mean the eventual total disappearance of the minority class, as models exist of stable trade-offs between minority and majority categories in language (Cuskley et al. 2014).

Most accounts of morphological levelling make one core assumption on the structure of the mental lexicon, the repository of words and constructions stored by the individual (Goldberg 1995). This assumption is that the mental lexicon is rich in detail, similar to broader, non-domain-specific cognitive category systems (Rácz et al. 2015; Goldinger 1997; Nosofsky 1988). As a consequence, words are organised according to similarity and frequency or predictability of use affects the strength of the individual representations.

A rich-lexicon account of morphological levelling, then, rests on the principles of **frequency** and **prototypicality**. Here, both frequency and prototypicality are broader characteristics of cognitive category organisation that are applicable to the organisation of the human mental lexicon.

Frequency has two aspects. First, word types with higher **token frequency** are more prone to morphological suppletion and tend to resist replacement and morphological levelling (Pagel et al. 2007; Albright & Hayes 2003). This can be easily seen in the case of the copula, which is one of the most frequent verbs and which shows rampant irregularity in almost every language. Second, word categories with higher **type frequency** tend to be more productive and to expand over time (Baayen 1993). Within morphologically complex forms, the relative frequency of the constituent parts has a huge effect on how the form is processed (Hay 2001). In a morphological levelling process, we expect more frequent types to resist levelling more.

Prototypicality has many working definitions, but it essentially relates to the **overall similarity** of a form to a given category (Gergely & Pléh 1994; Nosofsky 1988). A form is said to be prototypical if it is very close to the centre of the category. We can quantify this distance in various ways. In terms of linguistic categories, prototypicality relates to the extent to which a given word form is **similar** to other forms in the same category. In a morphological levelling process, we expect word types that have higher token frequency and that are closer to the centre of the minority category to be more entrenched in the category and in the lexicon and, as a consequence, to resist levelling more.

In the light of all this, we have to answer the following questions. Is variance in the 1SG.INDEF of the (-ik) class an instance of morphological levelling? If it is, how are the above principles manifest in this variation? Do we expect the (-ik) class to be completely levelled in with the regular class? How does social salience affect this process? In order to answer these questions, we first need to specify how the above, general, aspects of lexical organisation and morphological levelling apply to (-ik) verbs.

First, we ought to note that levelling should not be confined to variation in the 1SG.INDEF. We would also expect verbs to lose the (-ik) ending in the 3SG.INDEF. This is complicated by two factors. First, a large number of -∅/(-ik) pairs are related but distinct in meaning (as in the above [tør]/[tørik] example). This semantic distinction can serve to maintain a formal distinction.

Second, a number of (-ik) verbs would be illicit unsuffixed free forms in Hungarian in the 3SG.INDEF without the (-ik) suffix: [ɲugsik]/[*ɲugs] 'calm-INTR.3SG.INDEF'. (Compare with [raːg-s] 'chew-2SG.INDEF'.) I revisit these issues in section 6.

Focussing on forms and the 1SG.INDEF, we arrive at the following prediction:

(i) Verb forms that are attested with both (-ik) and -∅ forms will be more likely to use the **unmarked** 1SG.INDEF.

If we take a form that has (-ik) in the 3SG.INDEF, we expect it to prefer the **marked** 1SG.INDEF overall: [ɛsik] → [ɛsɛm] 'eat' 3SG/1SG.INDEF, [lɒkik] → [lɒkom] 'stay' 3SG/1SG.INDEF. Of course we expect to find unmarked forms as well, such as [ɛsɛk], [lɒkok].

If the verb has no (-ik) in the 3SG.INDEF, we expect the absence of a marked 1SG.INDEF: [ʃeːtaːl] ↛ [*ʃeːtaːlom] 'walk' 3SG/1SG.INDEF. If a verb is attested with both the (-ik) and the -∅ suffix, we expect it to be more likely to occur with the **unmarked** form in the 1SG.INDEF. ([bujdoʃ]/

[bujdoʃik] 'keep on hiding' → [bujdoʃok]). This is because these verbs are lexically less prototypical members of the (-ik) class.

Pairs of closely related, but semantically distinct verbs would confound this analysis. However, there are no such pairs in our data (such as [tør]/[tørik]).

Membership in both the regular and the (-ik) class should also affect prototypicality for **suffixes**. As we will see, many of the (-ik) verbs end in productive derivational suffixes, such as **deadjectival** (-odik)/(-ɛdik)/(-ødik) or **denominal** (-zik). We expect that if a suffix is mostly attested with regular verbs versus **-ik verbs**, this should manifest in a preference for the **unmarked** 1SG.INDEF (-ik) suffix – and **vice versa**:

(ii) a. Derived verbs in the (-ik) class will be more likely to use the marked 1SG.INDEF if the suffix itself has an overall preference for (-ik) in the 3SG.INDEF.

b. Derived verbs in the (-ik) class will be more likely to use the unmarked 1SG.INDEF if the suffix itself has an overall preference for -∅ in the 3SG.INDEF.

In sum, the degree to which both the verb and its constituent parts are exclusive to the minority (-ik) class should have a strong effect on suffix preference in 1SG.INDEF. The more the verb is closer to the centre of the (-ik) class, the more it should take the marked (-m) suffix that is characteristic of this class.

The general behaviour of the suffixes – whether they occur with (-ik) or -∅ across forms – will be probably relevant. It can be offset by properties of the verb stem, mainly its token frequency.

(iii) More frequent verbs in the (-ik) class will be more likely to use the marked 1SG.INDEF.

This is because a frequent form should, in models of the richly detailed mental lexicon, enjoy more autonomy and should be more resistant to levelling pressure. Given the inverse correlation between word length and word frequency (Zipf 1935), we expect longer forms – that are less frequent – to be less prone to use the marked form.

What other predictions can we draw from the interaction of frequency and prototypicality? Given that the marked 1SG.INDEF neutralises **definiteness** marking in the 1SG, we might expect (-ik) verbs that are **transitive** – and therefore have attested definite forms – to avoid using the marked indefinite in order to maintain expressivity:

(iv) If a verb in the (-ik) class has attested definite forms, it will avoid the marked indefinite form.

These numbered predictions follow if we regard variation in the 1SG.INDEF of the (-ik) stems in Hungarian as a symptom of a morphological levelling process. They are straightforward to test on a representative sample of the ambient language (Rácz et al. 2016). I will use the Hungarian Webcorpus to explore the degree to which they define variation in the (-ik) class.

2. The distribution of (-ik) variation

Data were collected from the frequency dictionary of the Hungarian Webcorpus (Trón et al. 2006). The dictionary is based on a version of the corpus that is morphologically analyzed (Trón et al. 2005) and morphologically disambiguated on the inflection level (Halácsy et al. 2007).

The Webcorpus (specified as above) contains 3887 types and 2,549,651 tokens of (-ik) verbs. 3121 types have attested 1SG.INDEF forms (164,627 tokens). I use an arbitrary cutoff rate to include verb types that have at least *four* 1SG.INDEF forms in total. These verb types all show variation in the 1SG.INDEF: they have at least one (-m) (marked) and one (-k) (unmarked) variant.

The size of this dataset is inflated by forms that are likely results of hypercorrect use. I consider [va:lom] 'transform-1SG.INDEF' and [ɛlva:lom] 'divorce-1SG.INDEF' ungrammatical, but they are both attested in the webcorpus (with a token frequency of 3 and 1, respectively).

My analysis of the data largely follows Janda et al.'s (2010) study of variation in Russian. I use the *R* open-source statistical environment for data analysis (R Development Core Team 2016) and the graphics package *ggplot* for visualisation (Wickham 2009). For each verb stem, I calculate the odds ratio of the marked and the unmarked 1SG.INDEF form. In order to make it easier to visualise, analyse, and interpret the data, I take the logarithm of the odds.

For instance, the verb [bulizik] 'party-3SG' has 48 attested 1SG.INDEF forms in the data, out of which 17 are marked and 31 are unmarked ([bulizom]/[bulizok]). The odds ratio of marked/unmarked for this verb is $17/31 = 0.55$. The log odds is -0.6.

The distribution of the log odds for all verbs in the dataset can be seen in Figure 1. The mean of the distribution is 0.96. We can see that the distribution is swerved towards the right – the positive values on the x axis indicate that, overall, the marked forms of the 1SG.INDEF are more frequent

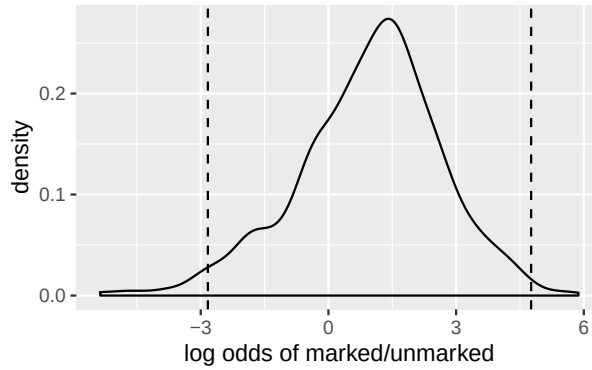

Figure 1: Density plot of the distribution of the variance across the 1SG.INDEF of the (-ik) verbs

than the unmarked ones. (In total, the dataset contains 125551 marked forms and 17128 unmarked forms.) This is in line with the descriptive assumption that (-ik) verbs prefer (-m) in the 1SG.INDEF. The preference is also at least partly due to the written domain that the corpus is sampled from. The dashed lines indicate 2.5 median absolute deviations from the mean. 24 verb types fall outside these cutoffs.

3. Operationalising our predictors

We now revisit the predictions of section 1, repeated below.

(i) Verb forms that are attested with both (-ik) and -∅ forms will be more likely to use the **unmarked** 1SG.INDEF.

(ii) a. Derived verbs in the (-ik) class will be more likely to use the marked 1SG.INDEF if the suffix itself has an overall preference for (-ik) in the 3SG.INDEF.
b. Derived verbs in the (-ik) class will be more likely to use the unmarked 1SG.INDEF if the suffix itself has an overall preference for ∅ in the 3SG.INDEF.

(iii) More frequent verbs in the (-ik) class will be more likely to use the marked 1SG.INDEF.

(iv) If a verb in the (-ik) class has attested definite forms, it will avoid the marked indefinite form.

How can we test for these predictions? Given the fuzzyness of existing semantic criteria on argument structure, I opt for a form-based approach, categorising verbs according to overlap in the shape of the stem.

For (i), I label the (-ik) verbs that are also attested without (-ik) in the 3SG.INDEF in the Webcorpus: 43 out of 825 verbs in the dataset. One example is [lɒviːrozik] 'manoeuvre-3SG.INDEF', also attested as [lɒviːroz]. I operationalise the effect of these forms not by their frequency in the webcorpus, but rather by whether they are present or absent.

For (ii), I focus on the suffix class defined as **medial derivational suffixes** by Abaffy (1978), viz., (-odik), (-oːdik), (-ɛdik), (-ødik), (-øːdik), as well as (-doklik), (-dɛklik), (-døklik), (-l(ik)), (-ll(ik)). My inspection of the 825 variable (-ik) stems reveals a third frequent derivational suffix, the **denominal** (-zik).

I group these variants into three classes, fully productive (-Odik) ([o/oː/ø/øːdik]) more lexicalised (-lik), as well as (-zik). I mark membership in these three classes for all verbs in my dataset. The single criterion of membership is formal similarity – the 3SG.INDEF of the verb has to end in (-Odik), (-lik), or (-zik).

For each suffix (the term *ending* might be more accurate, given the primacy of formal similarity in defining these classes), I count the number of **types** in the Hungarian Webcorpus, **with or without -ik**: (-od), (-oːd), (-ød), (-øːd) versus (-odik), (-oːdik), (-ødik), (-øːdik), (-øːdik); (-l versus (-lik), and (-z) versus (-zik).

I do not consider all possible variants.

First, **deadjectival** (-Odik) has a fifth variant, front unrounded (-ɛdik). This latter one, **without -ik**, overlaps with a separate verbal ending, (-Ad). As we will see, this is a relatively frequent form, so that (-ɛdik) is an exception to the overall pattern of (-Odik). This is excluded. We ought to also note that (-Odik) and (-zik) are fully productive, while (-lik) is not.

Second, I also exclude counts of (-ll) and (-llik). The former is a very rare verb ending, and so should not be compared to the latter. In order to control for this, only counts of (-l) and (-lik) are included. The summary can be seen in Table 2 below.

Subsuming (-odik), (-oːdik), (-ødik), (-øːdik) under the shorthand (-Odik) is idiosyncratic to this paper, as capital letter notations in morphophonology usually refer to allomorphs of the same morpheme. Short- and long-vowel variants of (-Odik) can also be analysed as separate suffixes (see Kiefer 2000).

Given that pairs of stems with (-ik) and -∅ (like [tør]/[tørik]) exist, a speaker can re-analyse -ed(ik), -Od(ik), -l(ik), and -z(ik) as pairs.

Table 2: Number of derived types with and without (-ik) in the Hungarian Webcorpus

Suffix	-ik	-∅
-ed(ik)	508	180
-Od(ik)	1853	18
-l(ik)	266	3684
-z(ik)	1131	1048

That is, one might assume that [jɛlɛz 'signify-3SG.INDEF' and [utɒzik] 'travel-3SG.INDEF' share a verbal derivational suffix, except that the latter is also an (-ik) stem. Then, a speaker has strong evidence that (-Od)/(-Odik) is a more typical (-ik) suffix, whereas [-l]/[-lik] is a more typical -∅ suffix. The presumed suffix (-z)/(-zik) has strong support for existing both as an (-ik) and a -∅ suffix.

Out of the 825 verbs, 23 end in (-lik), 344 end in (-zik), and 248 end in (-Odik).

For (iii), I used the logged lemma frequency of the verb as a proxy of the verb's overall frequency/predictability. Lemma frequency is the summed frequency of all attested forms of the verb in the corpus.

For (iv), I labelled the (-ik) verbs that have attested 1SG.DEF forms as well. That is, a verb is **transitive** if it has a **transitive** form in the corpus.

This operationalisation allows me to test for the prototypicality- and frequency-based predictions of treating variation in the (-ik) 1SG.INDEF as a case of morphological levelling. I almost entirely discarded semantics in favour of formal similarity and frequency of occurrence in the webcorpus. I see this as a feature, rather than a bug; definitions based on form are less ambiguous, and whether a verb has an **attested** transitive form is a clearer criterion of transitivity than whether the verb **could** have a semantic basis to be transitive.

4. The structure of (-ik) variation

I fit a mixed-effects logistic regression model on the dataset using the *lme4* package in R (Bates et al. 2015). The model predicts the log odds ratio of the marked (-m) over the unmarked (-k) variant of the 1SG.INDEF as a function of various characteristics of the verb form. The counts of the marked and the unmarked forms in the dataset are not independent – they are grouped under verb forms. (Each verb form has n marked and

m unmarked realisations.) To account for this lack of independence, the model also contains a random intercept for verb form.

The model's predictors are based on the indicators of morphological levelling outlined in the previous section. These are (a) the log lemma frequency of the verb form, (b) whether the verb has an attested 3SG.INDEF variant **without** (-ik), (c) whether the verb has an attested **transitive** form, (d) whether the verb ends in (-Odik) – ignoring [ɛdik] –, or (e) (-lik), or (f) (-zik). The summary of the fixed effects can be seen in Table 3.

When we look at the fixed effects, the estimate informs us of the relationship between predictor and marked/unmarked 1SG.INDEF variation across forms in the sample. The standard error of the estimate is a measure of accuracy. If we kept drawing samples from, say, online Hungarian use, the error gives us a sense of the variation of the estimate across these samples. If the estimate moves around a lot in repeated samples, it is not very informative of the real relationship in the population (e.g., variation in online Hungarian). Using a set of assumptions on the sampling distribution, we could calculate a *p* value to express the likelihood of the null hypothesis being true for each predictor, given the sample.

Since this analysis is highly exploratory and since, overall, there is no meaningful null hypothesis to reject, I will report the estimates and the standard errors for each predictor, but will not provide a *p* value. Instead, the Wald 95% confidence intervals are provided (see also Figure 2). If the intervals do not contain zero, we can be 95% certain that the effect is non-zero.

I go through the predictors one by one below and plot the predicted values for each predictor below using the *sjPlot* package in R (Lüdecke 2018).

Table 3: Estimated effects, standard errors, and Wald confidence intervals, logistic model

	Term	Estimate	Std. error	Statistic	2.5%	97.5%
1	(Intercept)	0.15	0.29	0.54	-0.41	0.71
2	log_verb_lemma_frequency	0.20	0.04	5.63	0.13	0.27
3	verb_is_transitiveTRUE	0.06	0.17	0.36	−0.28	0.40
4	verb_has_varying_ikTRUE	−1.12	0.26	−4.32	−1.63	−0.61
5	odik_suffixTRUE	0.42	0.16	2.68	0.11	0.73
6	lik_suffixTRUE	−1.80	0.37	−4.89	−2.53	−1.08
7	zik_suffixTRUE	0.16	0.14	1.10	−0.12	0.44
8	word_length	0.19	0.07	2.83	0.06	0.33

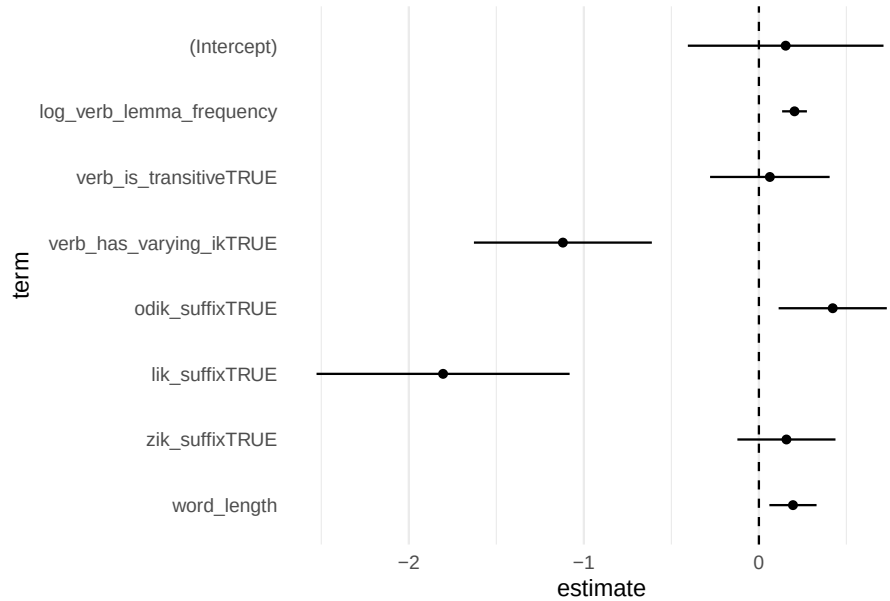

Figure 2: Estimated Wald 95% confidence intervals, logistic model

4.1. Verb frequency

The predicted effect of verb frequency can be seen in Figure 3. The positive correlation between frequency and preference for the marked form is very clear – while the marked form is preferred overall, this is especially true for frequent verb forms. This is congruent with the levelling hypothesis, since it shows that more frequent forms, those with a stronger lexical representation, are more resistant to adopting the majority variant (which is unmarked (-k) for **all** verbs).

Longer verbs **also** prefer the marked form (Figure 4 – note that syllable count was modelled as a continuouos variable to reduce model complexity). This remains true even though longer verbs tend to be less frequent. I would argue that this is a stylistic aspect of this variation. Longer verbs tend to be more formal ([ɛleːgɛdɛtlɛnkɛdik] 'dissent-3SG.INDEF', [ɛlbizoɲtɒlɒnodik] 'unsettle-3SG.INDEF'), which, in turn, leads to a more prevalent use of the formal, marked suffix.

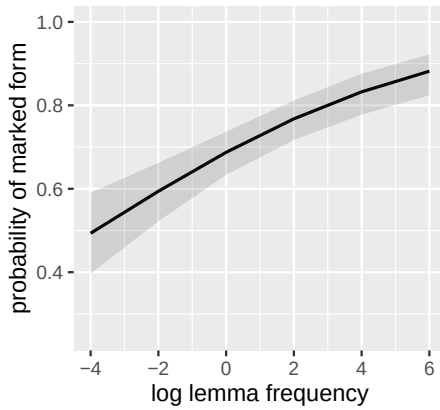

Figure 3: Model predictions: the effect of lemma frequency on marked/unmarked preference

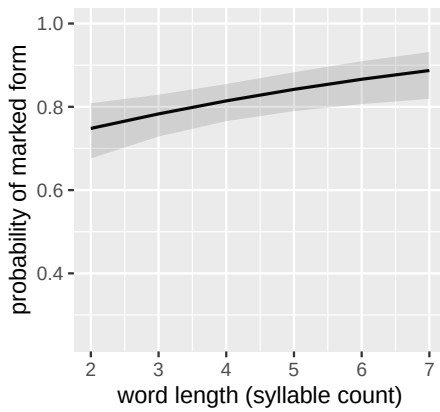

Figure 4: Model predictions: the effect of word length on marked/unmarked preference

4.2. Verb prototypicality

Those (-ik) verbs that have a -∅ variant in the 3SG.INDEF are far less likely to select the marked variant (Figure 5). This is, again, in line with the levelling hypothesis. If a verb shows variation in the – more stable – 3SG.INDEF, it is little suprise that it will be more prone to go for the majority unmarked pattern in the 1SG.INDEF.

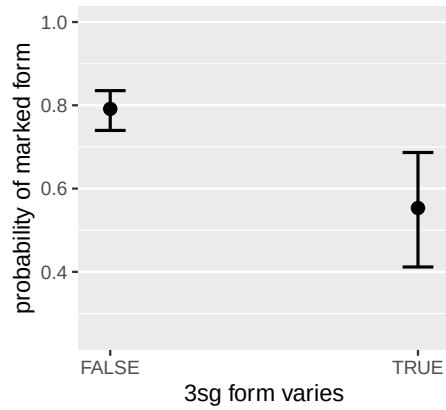

Figure 5: Model predictions: the effect of 3sg variance

Interestingly, whether a verb has an attested transitive form has little effect on its preference in the 1SG.INDEF (Figure 6).

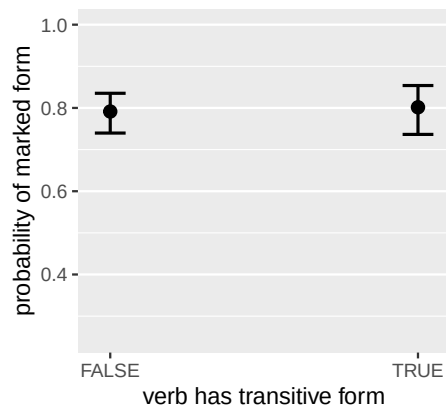

Figure 6: Model predictions: the effect of verb transitivity

This suggests that the system is not overly concerned with maintaining the **definiteness** distinction, at least not in the (-ik) class.

As seen in Figure 5, whether the verb is more centrally in the (-ik) class seems to matter. What about derivational suffixes? Recall that the distributions of our three derivational suffixes in the 3SG.INDEF provide different types of evidence on the degree **"ik-ness"** of the verb form. Most

(-Od)/(-Odik) forms are in the (-ik) class. Most (-l)/(-lik) forms are in the -∅ class. There is no such clear asymmetry for the (-z)/(-zik) forms.

The behaviour of the 1SG.INDEF of the derived forms follows precisely this pattern. For (-Odik) (Figure 7), where the ending is strongly suggestive of an (-ik) stem, we find a preference for the marked 1SG.INDEF. For (-lik), where the ending is strongly suggestive of a **regular** stem, we find the exact opposite (Figure 8). For (-zik), where the ending is not particularly informative of the verb class, we find no strong difference (Figure 9).

We should note that (-Odik), in particular, is a robust indicator of an (-ik) stem. This means that it is particularly easy to associate it with the 'correct' (socially accepted) marked suffix, even without invoking the stem itself. If a 1SG.INDEF ends in e.g., [-odo], the speaker will have a strong clue to conclude it to (-odom). Note, however, that (-ɛdik) constitutes an exception to this pattern. If we refit the model and include the [ɛ]-variant in (-Odik), the overall effect of this predictor is no longer robust. This makes sense if we consider the overlap between (-Odik) and (-Ad), discussed in section 3. Another thing to note is that the strength of (-Odik) as a predictor is weaker because the suffix itself is not very frequent in the sample.

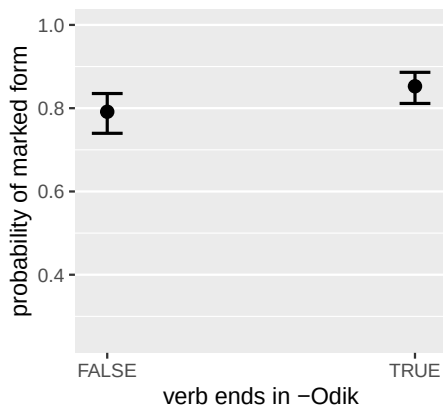

Figure 7: Model predictions: the effect of -Odik

These results are strongly indicative of a morphological levelling scenario in which verbs gradually move away from the (-ik) class towards the majority class. More frequent verbs, verbs that only have (-ik) forms in the 3SG.INDEF, and verbs that have the characteristic (-Odik) suffix, are more resistant to levelling. Verbs that have the characteristically **not** (-ik) suffix

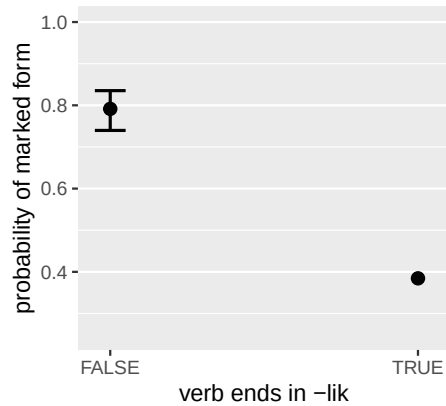

Figure 8: Model predictions: the effect of -lik

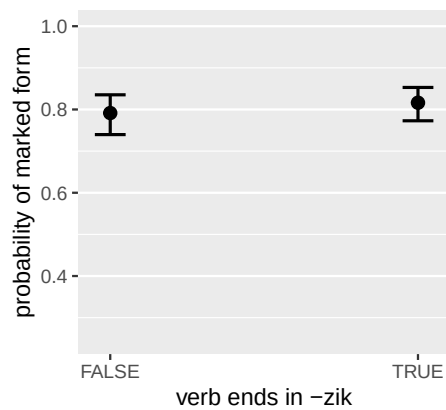

Figure 9: Model predictions: the effect of -zik

(-lik) are **less** likely to resist levelling. The marked 1SG.INDEF suffix of the (-ik) class neutralises the definiteness contrast, but this does not seem to be a relevant factor – (-ik) verbs that have transitive forms do not avoid the marked suffix to a larger degree.

5. Further sources of (-ik) variation

It is clear that the traits above are not the only ones responsible for variation in the 1SG.INDEF of (-ik) stems. The model residuals are indicative of further sources of variation.

The regression model predicts the log odds ratio of the marked and the unmarked variants of the 1SG.INDEF. Above, we discussed the estimated effect of our predictors – lemma frequency, suffix type, and so forth. What is also interesting is residual variation that is **not** explained by the regression model. Verb forms with high residual values are forms that the model 'missed', forms that are not behaving in a certain way because of the traits I used as predictors.

While I do not want to discuss model residuals in great detail, I note two types of outlier forms (forms that have a strong preference for the marked/unmarked suffix, not explained by the regression model).

First, short, frequent, intransitive verbs like [vaːlik] 'transform-3SG.INDEF' or [tyːnik] 'appear-3SG.INDEF' have a strong preference for the **unmarked** form.

Second, verbs that show variation between (-d) and (-s) in their conjugation show marked preferences. The (-d)/(-s) stems are a specific subclass of (-ik) verbs. It is possible for (-d)/(-s) forms to co-exist in one paradigm slot, as in [vɛrɛkɛdik]/[vɛrɛksik] 'fight-3SG.INDEF' or [mɛlɛgɛdik]/[mɛlɛgsik] 'get warm-3SG.INDEF'. For other verbs, the (-d)/(-s) forms co-exist within the paradigm, but have a set distribution, as in [fɛkydnɛ] 'lie-3SG.COND.INDEF'/ [fɛksik] 'lie-3SG.INDEF'. This is because the conditional suffix (-AnA) goes with the (-d) variant.

Here, the relevant aspect of this variation is that (-sik) forms overwhelmingly prefer the **marked** suffix in the 1SG.INDEF, while (-dik) forms prefer the **unmarked** suffix. This despite the fact that (-Odik), overall, shows a preference for the marked suffix. This can be seen in Table 4, which reports the mean odds of the marked and the unmarked form for (i) all forms, (ii) forms ending in (-dik), (iii) forms ending in (-sik).

Table 4: Preferences of (-dik) and (-sik) in the data

Suffix	Mean odds of marked/unmarked
(-dik)	6.64
(-sik)	27.11
all forms	8.97

Variation in the 1SG.INDEF has a non-linguistic aspect affecting its distribution in the corpus – it is socially salient. This can explain why the marked suffix is overrepresented in general. The corpus consists of written language (albeit written online), written language is more formal, and the marked suffix is more socially accepted, especially in formal registers. This also explains why very long, formal verbs prefer the marked suffix.

6. Discussion

I have proposed that variation in the 1SG.INDEF of the Hungarian (-ik) stems can be modelled as a case of morphological levelling, wherein verbs in the less populous (-ik) class are migrating to the larger, more regular -∅ class. I proposed a set of criteria to judge the validity of the morphological levelling scenario; frequent verbs, as well as more prototypically (-ik) stems and more prototypically (-ik) derivational suffixes should resist change. I used a dataset of verbs in the Hungarian Webcorpus to assess these criteria. Results support a morphological levelling scenario. My model does not cover all mechanisms of levelling – as evidenced by the existence of outlier forms such as [vaːlik] 'transform-3SG.INDEF' – and it cannot capture the social aspects of (-ik) variation. It still remains the first large-scale corpus-based quantitative study of variation in the 1SG.INDEF of the (-ik) stems, one that provides statistical evidence for an ongoing levelling process.

The levelling scenario has an apparent weakness. If we are witnessing morphological levelling, why has the (-ik) class not disappeared entirely? First, stable systems of majority and minority morphological classes exist, the best known one being the English past tense system (Cuskley et al. 2014). While the irregular class of English verbs has been shrinking since Old English, it has also been taking in new forms, and the irregular use of some very frequent forms has been stable through the history of the language. The Hungarian (-ik) class is much larger than the English irregular class. It is large enough to recruit new members in the 3SG.INDEF, which would then eventually also take up the remaining defining characteristics of the paradigm, most notably the marked 1SG.INDEF suffix. Second, Hungarian verbs constitute a lexically **closed** class, so that new forms must be formed with a derivational suffix, many of which, as we have seen, prefer (-ik) (e.g., [hɛkːɛrkɛdik] 'hack systems-3SG.INDEF ∼ [hɛkːɛrkɛdɛm] 1SG.INDEF, [intɛrnɛtɛzik] 'browse the Internet-3SG.INDEF) ∼ [intɛrnɛtɛzɛm] 1SG.INDEF). Native speaker judgements on such forms will likely vary, but that is beyond the scope of this short paper.

Third, a major reason for the existence of (-ik) verbs is morphophonological. Many of these verbs contain consonant clusters that are illegal at word boundaries in Hungarian, meaning that, without the (-ik) suffix, the stems are not viable. This is evidenced in various vowel alternation processes which result in a consonant cluster in certain paradigm slots and an intervening vowel in others. More notable is the apparent failure of these processes to operate in certain cases – the rampant defectivity of the Hungarian (-ik) stems in the imperative (Lukács et al. 2010), which lack all forms with consonant-initial suffixes. Overall, of course, the phonotactic requirement does not explain the persistence of (-ik) for verbs that would have a licit phonotactic configuration without the suffix, like [vaːlik] 'transform-3SG.INDEF' – hypothetical [vaːl] is well-formed.

The conflicting forces of Hungarian morphophonology will warrant the survival of the 3SG.INDEF (-ik). In turn, the strong social stigma associated with the only other part of the paradigm where the (-ik) class is consistently marked, the 1SG.INDEF, will push forms toward the marked (-m) suffix. This is further reinforced by the fact that (-Odik) is an excellent marker of the (-ik) class in the 1SG.INDEF, allowing speakers to keep track of suffix preference. The gradual levelling of the 1SG.INDEF of Hungarian verbal morphology is kept at bay by morphophonological and paradigm effects on the one hand, and social salience on the other.

Materials

For data and code, visit 10.5281/zenodo.3476934.

References

Abaffy, Erzsébet E. 1978. A mediális igékről [On medial verbs]. Magyar Nyelv 74. 280–293.

Albright, Adam and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. Cognition 90. 119–161.

Baayen, R. Harald. 1993. On frequency, transparency and productivity. In G. E. Booij and J. van Marle (eds.) Yearbook of morphology 1992. Dordrecht: Kluwer. 181–208.

Bates, Douglas, Martin Maechler, Ben Bolker and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67. 1–48.

Bickel, Balthasar, Bernard Comrie and Martin Haspelmath. 2008. The Leipzig glossing rules. Conventions for interlinear morpheme by morpheme glosses. https://www.eva.mpg.de/lingua/resources/glossing-rules.php. Accessed: 01-11-2017.

Bybee, Joan L. 1985. Morphology. A study of the relation between meaning and form. Amsterdam & Philadelphia: John Benjamins.

Cuskley, Christine F., Martina Pugliese, Claudio Castellano, Francesca Colaiori, Vittorio Loreto and Francesca Tria. 2014. Internal and external dynamics in language: Evidence from verb regularity in a historical corpus of English. PloS one 9. e102882.

Gergely, György and Csaba Pléh. 1994. Lexical processing in an agglutinative language and the organization of the lexicon. Folia Linguistica 28. 175–204.

Goldberg, Adele E. 1995. Constructions: A construction grammar approach to argument structure. Chicago: The University of Chicago Press.

Goldinger, Stephen D. 1997. Words and voices: Perception and production in an episodic lexicon. In K. Johnson and J. W. Mullenix (eds.) Talker variability in speech processing. San Diego: Academic Press. 33–66.

Halácsy, Péter, András Kornai and Csaba Oravecz. 2007. HunPos – An open source trigram tagger. In S. Ananiadou (ed.) Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Companion Volume. Proceedings of the Demo and Poster Sessions. Prague: Association for Computational Linguistics. 209–212.

Hay, Jennifer B. 2001. Lexical frequency in morphology: Is everything relative? Linguistics 39. 1041–1070.

Janda, Laura A., Tore Nesset and R. Harald Baayen. 2010. Capturing correlational structure in Russian paradigms: A case study in logistic mixed-effects modeling. Corpus Linguistics and Linguistic Theory 6. 29–48.

Kiefer, Ferenc (ed.). 2000. Strukturális magyar nyelvtan 3. Morfológia [A structural grammar of Hungarian 3. Morphology]. Budapest: Akadémiai Kiadó.

Kiss, Jenő and Ferenc Pusztai (eds.). 2003. Magyar nyelvtörténet [The history of Hungarian]. Budapest: Osiris Kiadó.

Kontra, Miklós and Tamás Váradi. 1997. The Budapest Sociolinguistic Interview: Version 3. Budapest: Nemzetközi Hungarológiai Központ.

Lukács, Ágnes, Péter Rebrus and Miklós Törkenczy. 2010. Defective verbal paradigms in Hungarian – Description and experimental study. In M. Baerman, G. G. Corbett and D. Brown (eds.) Defective paradigms. Missing forms and whatthey tell us. Oxford: British Academy. 85–102.

Lüdecke, Daniel. 2018. sjPlot: Data visualization for statistics in social science. R package version 2.6.2. https://CRAN.R-project.org/package=sjPlot

Nosofsky, Robert M. 1988. Similarity, frequency, and category representations. Journal of Experimental Psychology: Learning, Memory, and Cognition 14. 54–65.

Pagel, Mark, Quentin Atkinson and Alfred Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. Science 449. 717–721.

R Development Core Team. 2016. R: A language and environment for statistical computing. Vienna: Foundation for Statistical Computing. http://www.R-project.org

Rácz, Péter, Viktória Papp and Jennifer B. Hay. 2016. Frequency and corpora. In A. Hippisley and G. Stump (eds.) The Cambridge handbook of morphology. Cambridge: Cambridge University Press. 685–709.

Rácz, Péter, Janet B. Pierrehumbert, Jennifer B. Hay and Viktória Papp. 2015. Morphological emergence. In B. MacWhinney and W. O'Grady (eds.) The handbook of language emergence. Malden, MA & Oxford: Wiley Blackwell. 123–146.

Siptár, Péter and Miklós Törkenczy. 2000. The phonology of Hungarian. Oxford: Oxford University Press.

Trón, Viktor, Péter Halácsy, Péter Rebrus, András Rung, Péter Vajda and Eszter Simon. 2006. Morphdb.hu: Hungarian lexical database and morphological grammar. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk and D. Tapias (eds.) Proceedings of 5th International Conference on Language Resources and Evaluation. Genoa: European Language Resources Association (ELRA). 1670–1673.

Trón, Viktor, László Németh, Péter Halácsy, András Kornai, György Gyepesi and Dániel Varga. 2005. Hunmorph: Open source word analysis. In M. Jansche (ed.) Proceedings of the ACL Workshop on Software. Stroudsburg, PA: Associaton for Computational Linguistics. 77–85.

Wickham, Hadley. 2009. ggplot2: Elegant graphics for data analysis. New York: Springer.

Zipf, George Kingsley. 1935. The psycho-biology of language. Boston, MA: Houghton Mifflin Harcourt.