Acta Linguistica Academica 71 (2024) 1–2, 2–17
DOI: 10.1556/2062.2023.00658

AKADÉMIAI KIADÓ

Variation in the 1sg.indef: More than you wanted to know

PÉTER RÁCZ[1*] and ÁGNES LUKÁCS[1,2]

[1] Department of Cognitive Science, Faculty of Natural Sciences, Budapest University of Technology and Economics, Hungary

[2] MTA-BME Momentum Language Acquisition Research Group, Eötvös Loránd Research Network (ELKH), Hungary

Received: January 26, 2023 • Revised manuscript received: September 26, 2023 • Accepted: October 5, 2023
Published online: April 9, 2024

ABSTRACT

The first person singular indefinite or non-definite of Hungarian verbs that end in *-ik* shows variation between the regular *-k* suffix and the *-m* suffix, used otherwise in the definite. This variation is systematic and subject to metalinguistic awareness. Our study relies on previous quantitative work, a frequency dictionary compiled from the new Hungarian Webcorpus, as well as a forced-choice elicitation experiment to assess the role of word frequency, word length, derivational endings, and across-form similarity in shaping this variation. We find that first person singular indefinite variation is largely defined by natural categories: verbs that look similar will also show a similar preference to *-k*/*-m*. This pattern is attested in the webcorpus as well as in participant responses in the elicitation task.

KEYWORDS

morphology, variation, Hungarian, corpus linguistics, psycholinguistics

1. BACKGROUND

In Hungarian, verbs have definite and indefinite or non-definite forms. The definite verb form is used with definite objects and the indefinite is used everywhere else (*él.ek* 'live.1SG.INDEF', i.e. 'I live', *szeret.ek valaki.t* 'love.1SG.INDEF someone', i.e. 'I love somebody' vs. *kinevet.em őt*

* Corresponding author. E-mail: racz.peter.marton@ttk.bme.hu

'laugh.1SG.DEF him/her', i.e. 'I laugh at him/her'). This paper is about so-called Hungarian *-ik* verbs – ending in *-ik* in the third person singular present indicative (3SG) (e.g. *esz.ik* 'eat.3SG', *isz.ik* 'drink.3SG', *alsz.ik* 'sleep.3SG', *lak.ik* 'stay.3SG'), which show variation and can pick the definite ending *-m* instead of indefinite *-k* in the first person singular indefinite (*eszem* 'eat.1SG.DEF'; *eszek/eszem* 'eat.1SG.INDEF').

The *-k/-m* variation is a bona fide example of variation subject to social awareness in the Hungarian speech community and a perennial topic of metalinguistic debates in stylistics and education (Kálmán 2010). The neutralising *-m* is the literary variant and the variation indexes social status, education, and register.

Despite the variation's sociolinguistic salience, its stochastic linguistic aspects have been largely left unexplored. Do all *-ik* verbs vary and to what extent? Is this variation present both in and across speakers? What are the phonological and morphological characteristics affecting variation in this class? One exception is Rácz (2019) who provided a quantitative analysis of *-m* and *-k* forms of *-ik* verbs based on the frequency dictionary of the Hungarian Webcorpus (Halácsy et al. 2004; Trón et al. 2006).

In the current paper, we further explore the nature of this variation by addressing the following questions.

1. Does the 1SG.INDEF show systematic within- and across-word variation?
2. Is this variation driven by the form and frequency of the relevant forms?
3. Does this variation translate to within-speaker variation (where speakers use the two variants to different degrees)?

We draw on corpus data from both the Hungarian Webcorpus and the recently compiled Hungarian Webcorpus 2 (Nemeskey 2020). Both are web-scraped corpora. The first corpus consists of circa 1.5 billion words in total and 0.6 billion words after spell-checking. The second corpus consists of circa 9 billion words. We complement corpus data with the results of a forced-choice elicitation experiment using Hungarian nonce verb prompts.

2. 1SG.INDEF IN THE WEBCORPORA

There are about 2300 *-ik* verb lemmata (that end in <<ik$>> in 3SG.INDEF) with 1SG exponents in the frequency dictionary of the first Hungarian Webcorpus. Rácz (2019) found 825 that showed variation in the 1SG.INDEF. Using regression analysis to identify the factors predicting variation, the paper observed the following patterns:

- More frequent verbs were more likely to select the *-m* form over the *-k* form (for frequent *dolgoz.ik* 'work.3SG.INDEF', *dolgoz.om* 'work.1SG.INDEF' is more frequent, for less frequent *szaglász.ik* 'smell.3SG.INDEF', *szaglász.ok* 'smell.1SG.INDEF' is)
- In a range of 2–7 syllables, longer verbs were more likely to select *-m* over *-k* (*híz.ok>híz.om*, i.e. 'I gain weight' and *elbizonytalanod.ok<elbizonytalanod.om*, i.e. 'I become uncertain')
- Verbs that were attested in the corpus both with and without the 3SG *-ik* were less likely to select *-m* over *-k* (*böngész.ek>böngész.em* for *böngész/böngész.ik* 'browse.3SG.INDEF', cf. *eszek<esz.em* for **esz/esz.ik* 'eat.3SG.INDEF')
- Verbs ending in the derivational ending *-lik* were less likely to select *-m* over *-k* (*csukl.ok>csukl.om* for *csukl.ik* 'hiccup.3SG.INDEF')

This was a first attempt to disentangle and quantify factors influencing variation in *-m* and *-k* forms. While it was informative, some methodological problems and the development of new corpora call for further explorations of the matter. Here we provide a partial critique of the 2019 paper and test its predictions on the second Hungarian Webcorpus.

2.1. Data analysis

We fit two Bayesian generalised linear mixed models on the count ratios of *-k* and *-m* forms for variable verbs, one in Webcorpus 1 and one in Webcorpus 2, following Janda, Nesset & Baayen (2010). The predictors were log lemma frequency (scaled), stem syllable count (as a numeric variable), and derivational suffix. These were the predictors used by Rácz (2019). We unpack these, along with the predictors we skipped in the current analysis, below.

The models used a binomial error distribution, a logit link function, and weakly informative normal priors. We wanted to look at the ratios of *-k*/*-m* variants across verbs and discard raw counts. To achieve this, we modelled count ratios with a grouping factor for verb lemmata. We fit the models using RStan (Stan Development Team 2019) in R (R Core Team 2021).

2.2. Results

Figure 1 shows the effect estimates from the models, fit on Webcorpus 1 (left) and Webcorpus 2 (right). A Bayesian regression model gives a distribution instead of a point estimate to express the strength of a given predictor. We will report the median of this distribution as our estimate, its mean absolute deviation as its error, along with the 95% credible intervals. If this interval excludes 0, we can be 95% confident that the predictor's effect is non-zero.

2.2.1. Word frequency. First, we found that more frequent forms are more likely to take the *-m* variant than less frequent forms in both corpora. This effect is smaller in the (much larger) second Webcorpus. There are 966 verbs with variable 1sg.indef forms in Webcorpus 1 and 822 such verbs in Webcorpus 2. 469 verbs overlap. Figure 2 shows the log odds of *-k* over *-m* in the 1SG.INDEF along with the log lemma frequencies for these forms.

Log odds is the natural logarithm of a number whose numerator is the number of *-k* forms and the denominator is the number of *-m* forms, as attested in the corpus. A very high value here means a large majority of *-k* forms while a very low value means a large majority of *-m* forms in the 1SG.INDEF. Zero means parity. The verb *ábrándoz.ik* 'daydream.3SG' has 894 1SG.INDEF forms in Webcorpus 2, distributed over 284 *ábrándozok* and 610 *ábrándozom* forms. The probability of the *-k* form is $284/(284 + 610) = 0.31$, the odds of *-k*/*-m* is $284/610 = 0.47$, the log odds is -0.76, reflecting a large majority of *-m* forms. The top row of figures in Figure 2 shows the relationship between log lemma frequency (horizontal axis) and log odds of *-k*/*-m* (vertical axis, with a probability scale also provided for convenience on the right) in the two corpora (left and right).

According to our models, more frequent verbs are less likely to pick *-k* and more likely to pick *-m* than less frequent verbs in both corpora, but this effect is weaker in the second, larger corpus. This can be seen in Figure 1.

Going back to Figure 2, the bottom row of figures is shown largely as a quality control measure: for the 470 variable verb forms attested in both corpora, the log frequencies of the

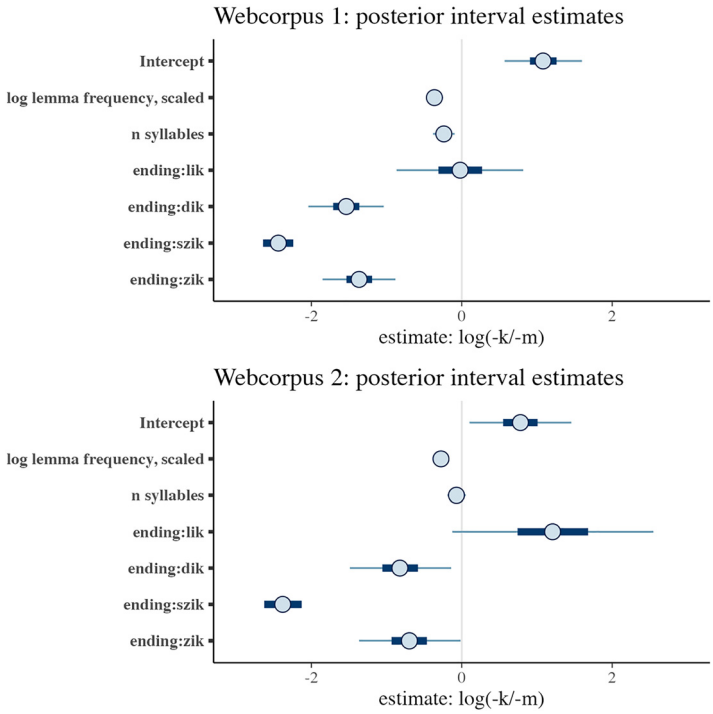

Figure 1. Posterior interval estimates from the model on Webcorpus 1 (top) and Webcorpus 2 (bottom). The circle marks the median of the posterior distribution for the predictor estimate, the thick band its 50% credible interval, its thin band its 95% credible interval. The vertical line is zero. If the thin band excludes the vertical line, the model is 95% sure that the real difference is non-zero. We now look at the results for the predictors in Rácz (2019) in turn. We will be referring back to this figure to assess the strength of the predictors.

individual lemmata in the two corpora are highly correlated. Overall, lemmata have lower raw log frequencies in Webcorpus 1 (x axis) than in Webcorpus 2 (x axis) because the latter is larger, but the relative frequencies pattern together.

2.2.2. Word length. Second, in both corpora, we reexamined the claim that longer verbs are more likely to select *-m* over *-k*. Figure 3 shows the distribution of log odds for *-k* over *-m* (vertical axis) across word length in syllables (specifically, the length of the stem in syllables, so that *lak.ik* 'live.3SG.INDEF' has a length of 1 and *aggód.ik* 'worry.3SG.INDEF' has a length of 2 on the horizontal axis).

We see an effect, present in the first Webcorpus, that is much less robust in the second, larger one (cf. Figure 1). While there is a steady trajectory of increasing preference for *-m* forms across length in Webcorpus 1, this trajectory is not straightforwardly present in Webcorpus 2. As a

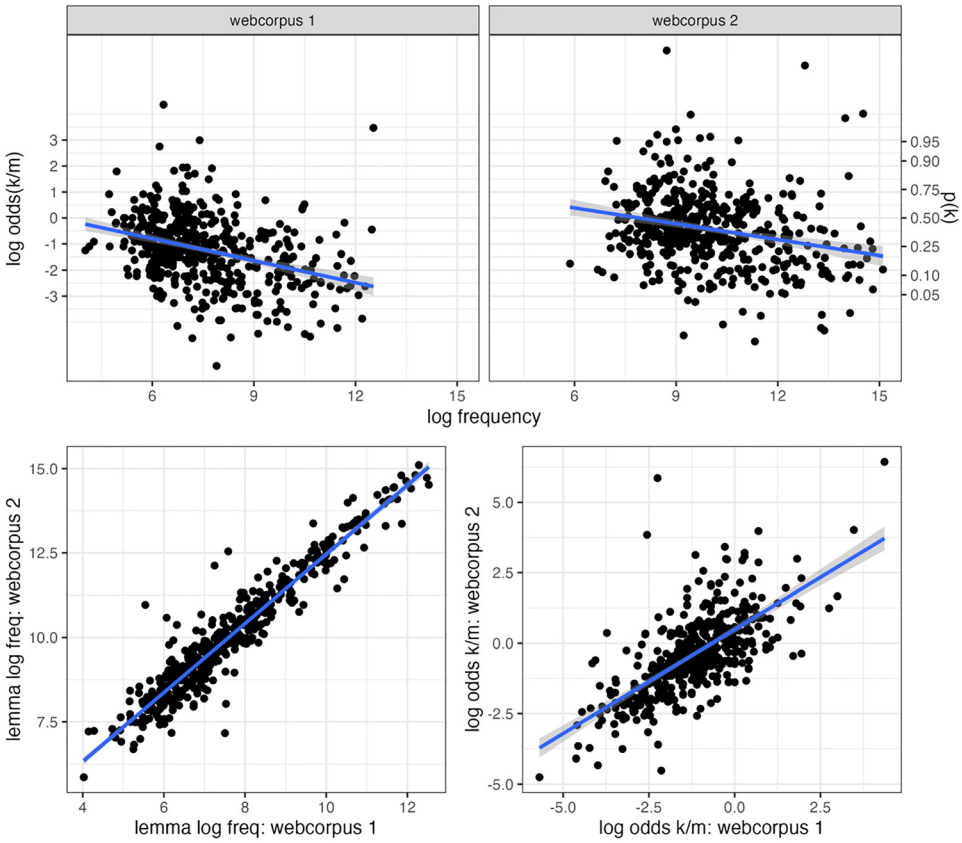

Figure 2. Frequency across log odds in the two webcorpora (top), Webcorpus 1 ∼ Webcorpus 2 for frequency (bottom left) and log odds (bottom right). Log odds express (logged) ratios of *-k/-m* per verb in the corpora. Log frequency is the log lemma frequency of the verb forms. A probability scale is provided for the reader's convenience on the right. Log odds are correlated with lemma frequency in both corpora (top) and log odds and frequencies are themselves correlated across corpora (bottom).

reviewer notes, length might be an epiphenomenon in the corpus data: longer words are more likely to have longer derivational suffixes that end in *-ik*, such as *-ódik*, *-kodik*, *-ódzik*.

2.2.3. Derivational endings. Unlike in English, verbs in Hungarian form a closed lexical class and a large number of new verbs are generated using a small set of derivational endings, including *-dik*, *-lik*, *-szik*, and *-zik*.

Calling these derivational endings is somewhat reductive, because the *-ik* variant is only present in the 3SG.INDEF. However, it can be seen as a variant of the derivational suffix that is present throughout the paradigm, e.g. *panír.oz.ok, panír.oz.ol, panír.oz/panír.oz.ik* ('breadcrumb 1-2-3SG.INDEF')

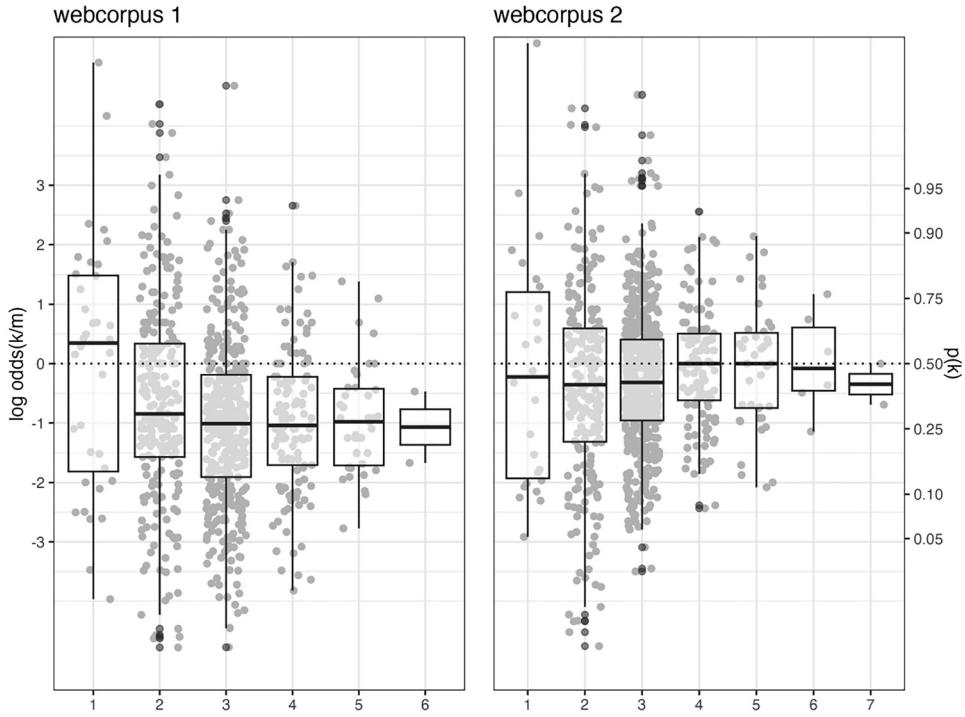

Figure 3. Log odds of *-k*/*-m* ratios across stem syllable count in the two corpora. A probability scale is provided on the right. Longer verbs are less likely to pick *-k* in the first corpus, this effect is not present in the second corpus.

Rácz (2019) finds that the relatively small set of intransitive verb forms that end in *-lik* strongly favour *-k* over *-m* (an exhaustive set fits in a pair of brackets: *botl.ik* 'stumble', *csukl.ik* 'hiccup', *dögl.ik* 'idle', *fényl.ik* 'shine', *haldokl.ik* 'die', *hanyatl.ik* 'decline', *oml.ik* 'collapse', *sikl.ik* 'slide', *sínyl.ik* 'suffer', *tündökl.ik* 'glare', *vonagl.ik* 'slither'). This pattern is also clearly visible in the larger Webcorpus 2, as seen in Figure 4.

What are all the relevant differences here? We can compare subgroups using non-linear hypothesis testing (Bürkner 2018). In Webcorpus 1, *-lik* verbs and verbs that have no derivational ending clearly favour *-k* over *-m*. Verbs that end in *-dik* or *-zik* pattern differently, favouring *-m* over *-k*. Verbs ending in *-szik* prefer *-m* even more. We see effectively the same patterns in Webcorpus 2, except that the three subgroups are more different and that *-dik* and *-zik* are much closer to *-k*/*-m* parity. This can be seen in Figure 1.

It is interesting to note that, in some way, *-lik* verbs are *-ik* verbs that are least likely to be confused with the majority non-*ik* class (Lukács, Rebrus & Törkenczy 2010): they are a closed class, never surface without the *-ik* endings in 3SG.INDEF (**csukol* 'hiccup.3SG'), and have no attested forms with suffixes that have no vowel-initial alternants (so-called analytic suffixes, see Siptár & Törkenczy 2000). One example is the imperative (**csukol.ja* 'hiccup.3SG.DEF',

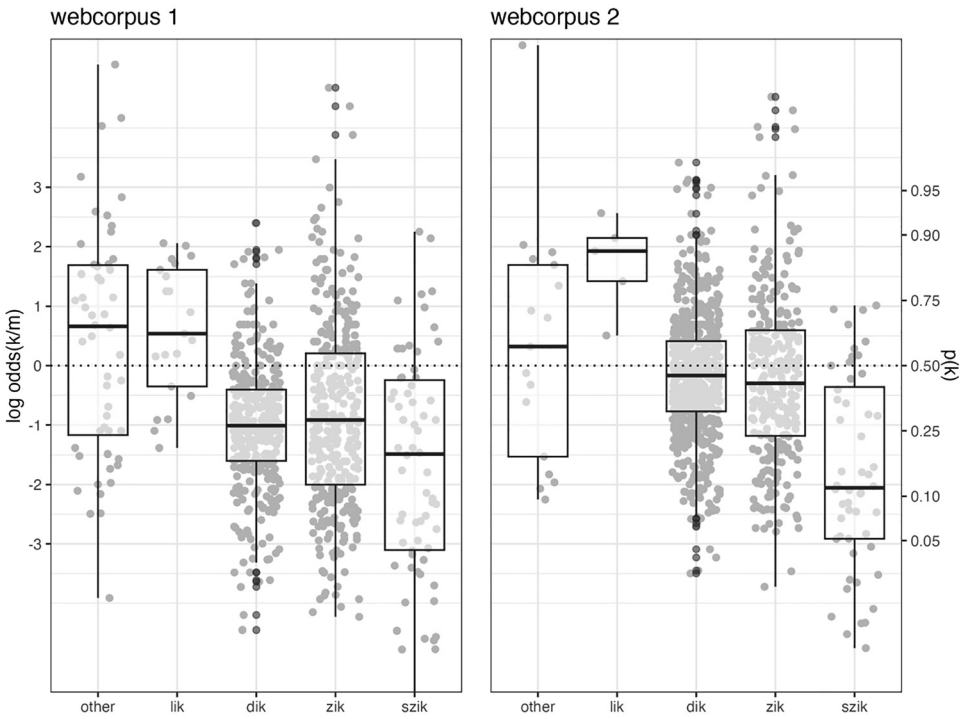

Figure 4. Log odds of *-k/-m* ratios across ending types in the two corpora. A probability scale is provided on the right. Forms that have no recognisable derivative ending (like *szűn.ik* 'cease.3SG.INDEF', *szök.ik* 'flee.3SG.INDEF') are labelled as "other". As a reviewer notes, *sínyl.ik* with *-ik* is obsolete/dialectal, and the colloquial form of this verb has no 3SG.INDEF; only the definite form is attested (*sínyl.i*).

**csukol.j* 'hiccup.1SG'), roughly analogous to Norwegian defectives (like *sykler* 'bicycle.3SG' / **sykl* 'bicycle.3SG.IMP').

An alternative explanation for the difference in distributions across corpora for *-lik* is that this might be a coding artefact: some verbs hesitate between the *-lik* and the *-l* form in their stems, and this is not discussed in Rácz (2019). This might indicate that Rácz (2019) excluded some variable forms that are otherwise more likely to behave like an *-ik* verb.

2.2.4. Variable ending and transitivity. Rácz (2019) looked at two more possible effects on 1SG.INDEF variation; whether the verb is attested without the 3SG *-ik* (*böngész/böngész.ik* 'browse.3SG') and whether the verb is transitive (*esz.ik* 'eat.3SG', *lak.ik* 'stay.3SG'). These cannot be tested based on a frequency list alone.

First, variation in *-ik* is syntactically motivated: verbs with a 1sg or 2SG object are more likely to drop the *-ik* ending: *A főnök jól megdolgoz engem* 'the boss well work.3SG.DEF me', i.e. 'My boss is driving me hard' vs. **János szaporán dolgoz.* 'John fast work.3SG.INDEF', i.e. 'John works fast'. A purely lexical analysis cannot filter for syntactic context and collapses this distinction (Rebrus,

Péter p.c.). In addition, one can even argue that the preverb and the verb create a new lemma (*megdolgoz*), which follows its own patterns of variation.

Second, Hungarian verbs are transitive or intransitive, but any verb (apart from the copula and a few motion verbs) can have a direct object, as in *Végigálmodoz.om a délutánt* 'daydream.1SG.DEF the afternoon', i.e. 'I daydream through the afternoon'. This means that we can label verbs in a frequency list as transitive (like *esz.ik* 'eat.3SG') or intransitive (like *lak.ik* 'stay.3SG') but this will not be informative on how much speakers use these forms with direct objects.

2.3. Discussion

Analysis of the two corpora reveals that there is a consistent effect of lemma frequency and ending type on the preference of *-k*/*-m*. The effect of word length is not present in the larger corpus.

Rácz (2019), largely on grounds of the frequency and length effects observed in Webcorpus 1, argues that the distribution of *-k* and *-m* forms is consistent with a morphological levelling scenario, in which the majority non-*ik* verb class absorbs the minority *-ik* class, pushing 1SG.INDEF forms towards the *-k* ending. This account says that there were more verbs behaving in a manner consistent with the *-ik* class, and their numbers have gradually diminished over time as verbs switched allegiance to the majority class. This is roughly analogous with the shrinking of the irregular verb class in English (where, over time, *wove* became *weaved*, *dove* became *dived*, etc).

Data from the new webcorpus do not unequivocally support the levelling account for Hungarian. Webcorpus 2 has more types and much more tokens than Webcorpus 1 and we should assign more reliability to its results. The fact that the frequency effect on variation is stronger in Webcorpus 1 is likely an artefact of corpus size: more frequent *-ik* verbs are more likely to be sampled into a smaller corpus. There is a strong native speaker intuition that there is a set of frequent forms which is especially unlikely to select the more common *-k* variant (Kálmán 2010). These will be overrepresented in the smaller corpus and we will see a frequency effect. The frequency effect is not entirely due to sampling, however. It is present in the larger corpus, and the relationship of frequency and preference for k/m is linear: within each corpus, more frequent forms are more likely to behave in a certain way than less frequent forms.

But we cannot say that this effect is present because the likelihood of the *-m* form gradually increases with frequency, for reasons of lexical strength. The same is true for word length – especially since word length and frequency are correlated in natural language corpora.

An alternative account could posit that *-ik* verbs constitute a stable class of variable forms and class membership is largely explained by formal characteristics: verbs that look a lot like other *-ik* verbs will behave like them. This is supported by the patterns of ending types in both corpora: There are verbs ending in *-ik* that are more characteristic in picking the marked *-m* variant in the 1SG.INDEF (like verbs ending in *-szik* and *-dik*) and there are verbs that are less so (those ending in *-lik* and *-zik*). These results raise the possibility that similarity to other forms has an influence on 1SG.INDEF variation.

This account suggests that a range of verb-specific factors influence 1SG.INDEF variation. This means that similarity to specific verbs will predict behaviour that is similar to how those specific

verbs behave in terms of selecting a 1SG.INDEF variant. The way to test this is to expose language users to nonce verbs that look like real verbs but have no prior distributions of 1SG.INDEF variation. We ran a forced-choice elicitation study with nonce verbs in the 1SG.INDEF to explore the effects of similarity.

3. ELICITATION EXPERIMENT

We wanted to see whether nonce verbs that look similar to existing *-ik* verbs would elicit the same variation (*-m* over *-k* in the 1SG.INDEF) that we observed in corpus data. We wanted to test this in a semantically and syntactically simple environment in which individual participants respond to nonce verbs in simple prompts, following Berko's WUG paradigm (Berko 1958).

3.1. Methods

3.1.1. Participants. 85 students of the Budapest University of Technology and Economics took up the task for course credit in the spring of 2022, 78 finished it (67 women, median age 22). The study was approved by the United Ethical Review Committee for Research in Psychology in Hungary (EPKEB, ref. number 2021-119).

3.1.2. Stimuli. We generated nonce verb forms with different degrees of similarity to existing verb forms. To establish similarity measures, we first defined a training set of existing forms by drawing all 3SG.INDEF verb forms ending in derivational *-lik*/*-szik*/*-zik*/*-dik* with a stem length of 1 or 2 syllables from a spellchecked frequency list compiled from Webcorpus 2 (171 forms). The rationale behind this was that new verbs are formed exclusively through derivational endings, either from loanwords (*gugliz.ik* 'google.3SG.INDEF', *squash.ol* 'play squash.3SG.INDEF') or existing forms (*depised.ik* 'get depressed.3SG.INDEF') and so nonce verbs with such endings would be more palatable to native speakers.

We parsed the forms into syllabic constituents and built weighted n-gram models to generate nonce verb forms with 1- and 2-syllable stems. We filtered the resulting lists to make sure all forms were at least at an edit distance of 1 (for 1-syllable stems) or 2 (for 2-syllable stems) from (i) each other and (ii) all words in the spelling dictionary, and (iii) had no initial or final overlaps with words where the overlap was longer than 3 characters in the spelling dictionary. The resulting list was hand-filtered.

-dik forms were absent from the final list, largely because these forms tend to have a linking vowel so that the minimum stem syllable count is 2. This is more a design-based reason than a language-based one. The final list consisted of 162 forms, 81 with a stem length of 1 syllable, 81 with a length of 2, 79 ending in *-lik*, 30 in *-szik*, and 53 in *-zik*.

3.1.3. Procedure. The task was coded in Psychopy (Peirce et al. 2019). Each participant completed the task on their home computer via Pavlovia, an online experimental platform (https://pavlovia.org/). Each participant responded to 162 prompts with a binary forced-choice response. 54 were *-k*/*-m* prompts, presented in random order. Participants were instructed that

they would see Hungarian words, which might not be familiar to them, and would have to pick a suitable form in the target sentence for the word in the prompt sentence. They were given an example before starting.

The prompt-response structure looked like this:

- Te bizony sokat pratánylasz. (you.sg sure often pratánylik.2SG.INDEF 'You sure pratánylik a lot')
- Én is sokat… ('I also often…')
- pratánylok / pratánylom (pratánylik.1SG.INDEF)

3.1.4. Data analysis. We fit a Bayesian generalised linear mixed model on the data with weakly informative normal priors, a binomial error distribution and a logit link function. The outcome variable was the *-k*/*-m* response given by participants to nonce word prompts. There were three predictors: nonce stem length (1 or 2 syllables), nonce stem ending (*-lik*/*-zik*/*-szik*), and formal similarity. Data were grouped across respondent and target word.

Formal similarity was calculated as follows. We took the list of verbs with variable 1SG.INDEF forms from Webcorpus 2, split the list into thirds based on the preference for *-k* over *-m*, considered the first third (relative preference for *-k*) as the k training set, the third third (relative preference for *-m*) as the m training set, and discarded the middle third.

We used the Generalised Context Model, a categorisation model Nosofsky (2011) adapted from Dawdy-Hesterberg & Pierrehumbert (2014) by Rácz, Rebrus & Törkenczy (2021). The Generalised Context Model takes a target form and calculates its formal similarity to members of two or more training sets. In this particular case, the targets were the 3SG.INDEF forms of the nonce verbs in the experiment, transcribed phonemically. The model calculated their similarity to the 3SG.INDEF forms of existing verbs in the two training sets, and returned two category weights for each target: one for m and one for k. These weights sum up to 1 and express the extent to which a nonce verb is similar to the k or the m set.

We tested whether the effect of formal similarity was different across the three derivational endings (*-lik*, *-szik, -zik*). We used leave-one-out cross-validation to compare the log pointwise predictive densities of a model with and without the interaction. We found that a model with the interaction does not provide a better fit of the data and report the model without the interaction.

3.2. Results

Participants and words vary in *-k*/*-m* preference in the task. This is driven by the nonce verbs' ending types, and, far more importantly, by their similarity to real verbs.

Figure 5 shows the distribution of participants (left) and words (right) across how much they preferred *-k* over *-m*. Most participants and words show variation between *-k*/*-m*. This suggests that 1SG.INDEF variation is present both across words (where some people will say *lak.ok* 'stay.1SG.INDEF', others *lak.om*) and across participants (where someone might sometimes say *lak.ok* or *lak.om*).

Figure 6 shows the posterior interval estimates from the model fit on the data.

Figure 7 shows the log odds of participant *-k*/*-m* responses across the three endings. Words that end in *-lik* (like *spáraklik*) are most likely to select *-k* (*spáraklok)*, those that end in *-szik* (*piragszik*) are least likely (so *piragszom*), with the *-zik* class (*drügzik*) in between. The *-szik*/*-zik*

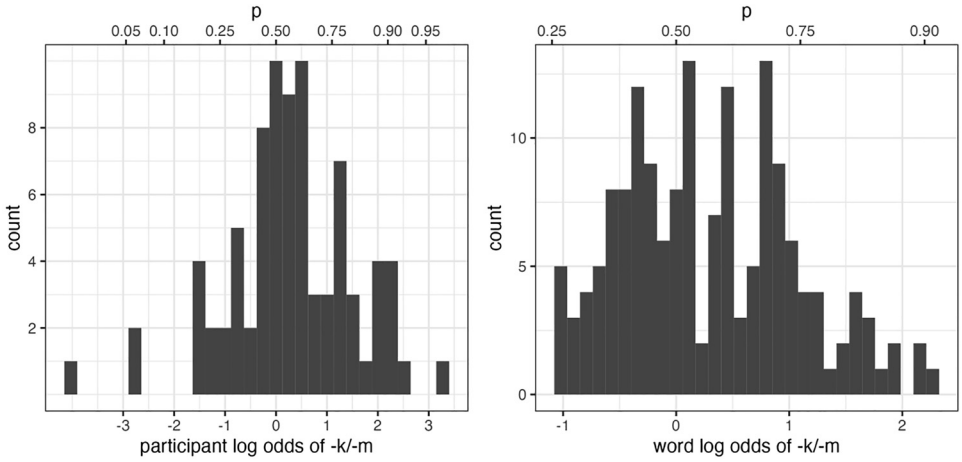

Figure 5. Participant and word distributions of *-k/-m* preference in the forced-choice elicitation task. A probability scale is provided on the top.

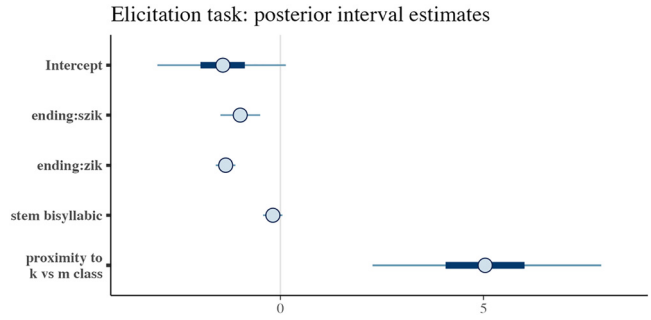

Figure 6. Posterior interval estimates from the model on the elicitation task. The horizontal axis shows the sampled values from the posterior for each predictor estimate. The circle marks the median of the posterior distribution, the thick band its 50% credible interval, its thin band its 95% credible interval. The vertical line is zero. If the thin band excludes the vertical line, the model is 95% sure that the real difference is non-zero.

difference is not statistically meaningful, however. How does this relate to real words? Here, we recover the *-lik*/*-szik* difference seen in the corpora. But the smaller *-szik*/*-zik* difference present in the corpora does not replicate in the elicited production results. (Words without a gloss are nonce forms.)

Figure 8 shows that verbs with bisyllabic stems (*málapszik*) are slightly more likely to select *-m* (*málapszom*) than those with monosyllabic stems (*jüslik, jüslök*). This is not a robust difference (see Figure 6).

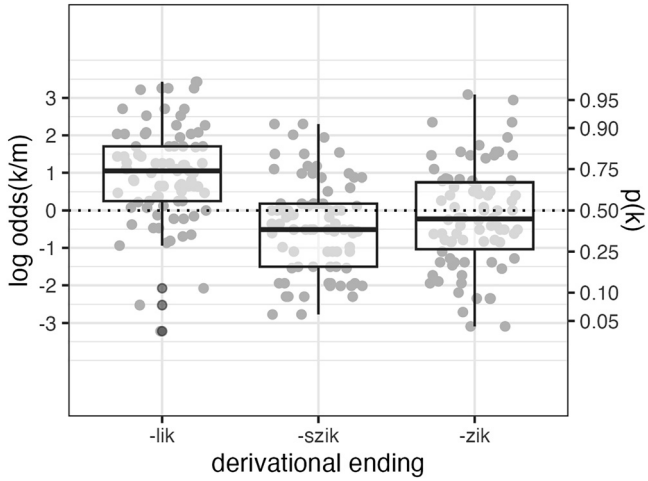

Figure 7. Log odds of *-k/-m* per participant across ending types in the forced-choice elicitation task. A probability scale is provided on the right.

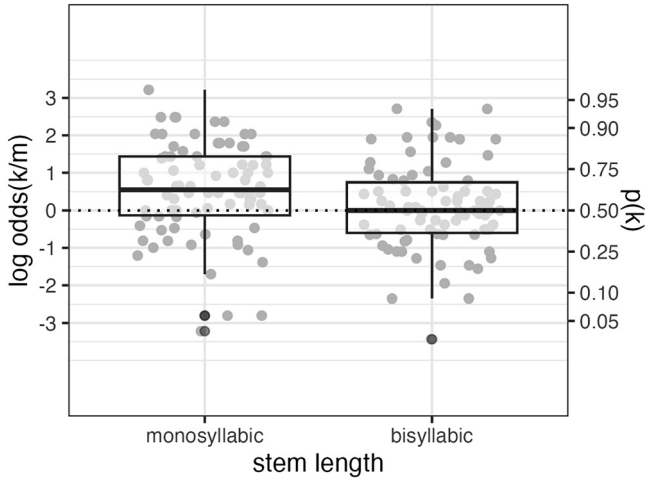

Figure 8. Log odds of *-k/-m* per participant across stem length in the forced-choice elicitation task. A probability scale is provided on the right.

Figure 9 shows log odds across model weights for individual words in the task. Words that look more like real words that prefer *-k* over *-m* also prefer *-k* over *-m*. This is by far the strongest predictor in the model, albeit also the noisiest (see Figure 6).

The generalised context model considers real verbs that strongly prefer the *-k* variant over the *-m* variant (like *bűnöz.ök* 'commit crimes.1SG.INDEF', *távoz.ok* 'depart.1SG.INDEF', *szűnök* 'cease.1SG.INDEF', *főzőcskéz.ek* 'cook.hab.1SG.INDEF', *pihiz.ek* 'rest.1SG.INDEF', *tartoz.ok* 'owe.1SG.INDEF') as one class and

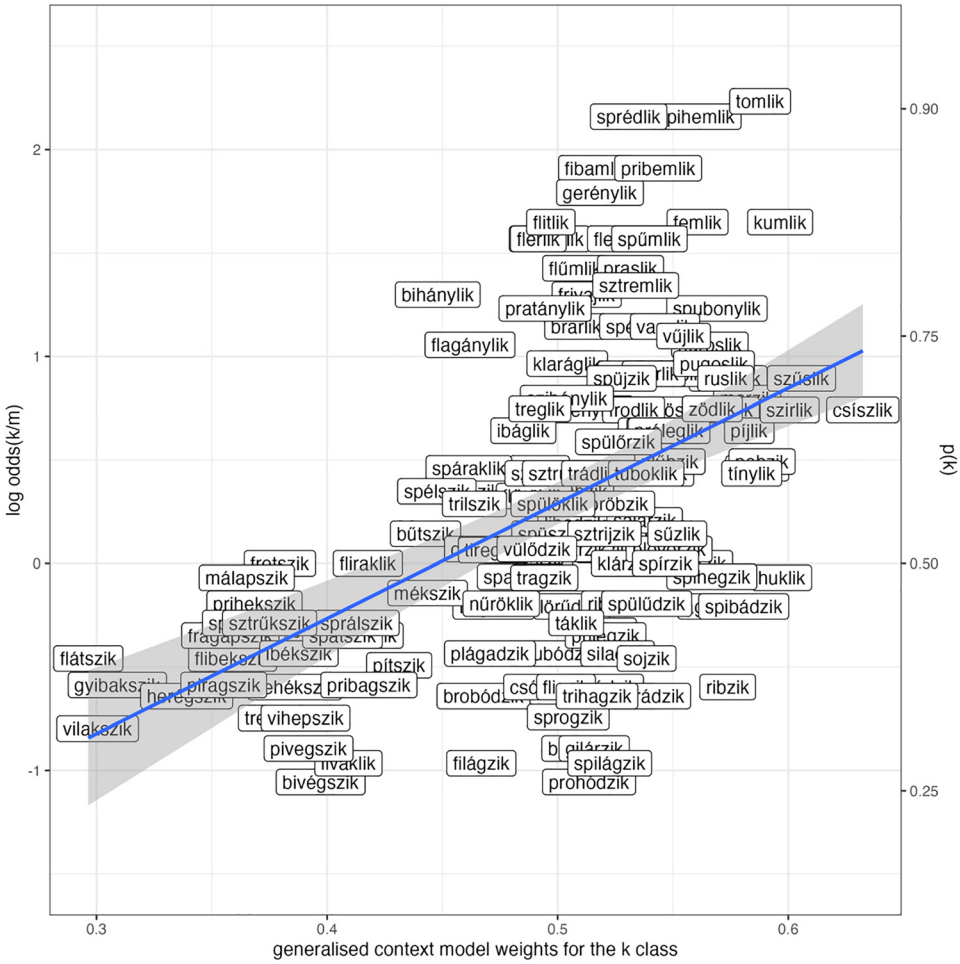

Figure 9. Model weights and log odds of *-k/-m* per word in the forced-choice elicitation task. A probability scale is provided on the right.

those that prefer the *-m* variant (like *esküsz.öm* 'swear.1SG.INDEF', *tolaksz.om* 'be inquisitive.1SG.INDEF', *alkusz.om* 'haggle.1SG.INDEF', *töreksz.em* 'thrive.1SG.INDEF', *emléksz.em* 'reminisce.1SG.INDEF', *öregsz.em* 'age.1SG.INDEF') as another. Then, it takes each nonce word (like *sprédlik, vilakszik, flátszik*) and calculates the distance between the word and these two classes. The two distances add up to one. So *vilakszik* has a distance of 0.27 from the k class and 0.73 from the m class. *Csíszlik* has a distance of 0.61 from the k class and 0.39 from the m class (see Figure 9). If a verb looks more like the k class than the m class, participants will be more likely to select the *-k* variant for it (*vilakszom, csíszlok*).

Overall similarity (including similarity to endings) to *-k* and *-m* forms and ending type have independent contributions to explaining variation in the experiment. This means that both are important: when participants need to determine whether to pick *spubonylok* or *spubonylom* as

the 3SG.INDEF of *spubonylik*, this decision seems to be shaped both by the target's similarity to existing *-k* verbs in general and the fact that it ends in *-lik* in particular. The two factors do not show an interaction: similarity is equally important across all endings.

As defined in our experiment, similarity between words includes similarity across their endings: two forms, both ending in *-lik,* are, in some way, more similar to one another than one ending in *-lik* and one in *-szik*. This means that *-k/m* similarity and ending type are two predictors that overlap. We can quantify this insight: the r^2 of a linear model predicting k weight from l ending type across words is 0.64, which is substantial. However, they both explain variance in our model, as seen in Figure 6. This means that a nonce verb's overall similarity to real verbs that prefer *-k* or *-m* is informative above and beyond whether this nonce verb ends in *-lik* or *-szik*. Conversely, whether the verb ends in one or the other is by and large more informative than the single edit distance of difference between the two endings.

3.3. Discussion

The replicated corpus analysis and the results of the forced-choice elicitation task indicate that 1SG.INDEF variation in Hungarian shows both within-and across-word and within- and across-speaker variation, and that this variation is largely determined by form and similarity (Figures 1 and 6), more so than frequency (Figure 1).

Rácz (2019) pointed to frequency and length effects to argue that the *-k*/*-m* variation in the 1SG.INDEF of the *-ik* verb class is undergoing morphological levelling. This paper revisits this idea using new corpus analyses and forced-choice elicitation data. In our corpus analysis, we find that existing verbs are less likely to prefer *-k* if they are more frequent overall and that verbs with similar ending types pattern together. In Section 2.2.3, we speculate on the morphophonological groundings of this pattern. In our elicitation task, we find that 1sg.indef *-m* is productive and the extent of its variability is driven by similarity: if nonce verbs look like real verbs that prefer the *-m* variant, these nonce verbs will also prefer the *-m* variant.

The available evidence does not support the levelling scenario outlined in Rácz (2019). Instead, it looks like *-ik* verbs form a continuum. The *-m* pattern is marked in the sense that it is the exponent of the 1SG.DEF and is not used in the 1SG.INDEF with verbs outside this class, apart from a few hypercorrect forms, such as *könyörög* 'implore.3SG.INDEF', *könyörgöm* 'implore.1SG.INDEF'. Some verbs are more likely to use this marked pattern than others.

Although frequent forms also tend to prefer the *-m* variant, it might have extraneous, social reasons that affect frequent forms more than infrequent ones. Hungarian 1SG.INDEF variation carries second-order indexicality (Eckert 2008), with the *-m* variants regarded as more elegant and educated. Metalinguistic awareness might be more restricted to specific, common forms, explaining why speakers use them more often.

Generally, similar high frequency forms constitute a lexical gang that exerts influence on neighbouring forms. For instance, English irregular verbs of the form *swim, sing, ring* create a somewhat productive ablaut pattern. If a form is too frequent, it will be emancipated and isolated in the lexicon, exerting little influence on any adjacent forms, as in the case of suppletive *go, be, have* in English (Bybee & Moder 1983). It is possible that *gyarapsz.ik* 'multiply.3SG', *veszeksz.ik* 'bicker.3SG' influence 1SG.INDEF variation but more frequent *esz.ik* 'eat.3SG', *lak.ik* 'stay.3SG' do not. It is also possible that the former are subject to analogical pressures in the lexicon (such that words that look like other words will behave like other words) and so will

show a preference for the *-m* variant, while the latter will be subject to conscious metalinguistic scrutiny and will be effectively rote-learned with the "formal" *-m* variant. These hypotheses are to be tested by further experiments with nonce verbs where similarity to specific forms with different frequencies is more directly controlled.

In this paper we have shown that frequency, ending type and word form similarity are important agents in determining morphological variation of 1SG.INDEF forms. The results also highlighted some of the problems of isolating the effects affecting variation: similarity interacts with syntactic, semantic and discourse effects, it is multidimensional, dimensions may overlap, they can be weighted differently in different contexts. It has become clear that these effects are often difficult to disentangle and can only offer limited insight. The emerging patterns and limitations underline the importance of a key point in Kálmán's thinking about language which keeps inspiring his colleagues and students: that variation in language is best considered using a holistic and analogy-based approach (Kálmán 2008).

Funding: This work was supported by Hungarian Academy of Sciences, Momentum 96233; Hungarian Scientific Research Fund, OTKA 138188; Hungarian National Research, Development and Innovation Fund, TKP2023.

ACKNOWLEDGEMENTS

The authors would like to thank their editor and reviewers.

REFERENCES

Berko, Jean. 1958. The child's learning of English morphology. Word 14. 150–177.

Bürkner, Paul-Christian. 2018. Advanced Bayesian multilevel modeling with the R package brms. The R Journal 10(1). 395–411. https://doi.org/10.32614/RJ-2018-017.

Bybee, Joan L. and Carol Lynn Moder. 1983. Morphological classes as natural categories. Language. 251–270.

Dawdy-Hesterberg, Lisa Garnand and Janet Breckenridge Pierrehumbert. 2014. Learnability and generalisation of Arabic broken plural nouns. Language, Cognition and Neuroscience 29(10). 1268–1282.

Eckert, Penelope. 2008. Variation and the indexical field. Journal of Sociolinguistics 12(4). 453–476.

Halácsy, Péter, András Kornai, László Németh, András Rung, István Szakadát and Viktor Trón. 2004. Creating open language resources for Hungarian. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004). 1201–1204.

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(1). 29–48.

Kálmán, László. 2008. The holistic view in linguistics. English translation of L. Kálmán (2007) Holisztikus szemlélet a nyelvészetben. Szabad Változók 4.

Kálmán, László. 2010. Néhány szó az ikes ragozásról. [A few words on ik verbs]. Nyelv és Tudomány, December 28, 2010. https://www.nyest.hu/hirek/nehany-szo-az-ikes-ragozasrol.

Lukács, Ágnes, Péter Rebrus and Miklós Törkenczy. 2010. Defective verbal paradigms in Hungarian—Description and experimental study. Proceedings of the British Academy 163. 85–102.

Nemeskey, Dávid Márk. 2020. Natural language processing methods for language modeling. Doctoral dissertation. Eötvös Loránd University, Budapest.

Nosofsky, Robert M. 2011. The generalized context model: An exemplar model of classification. In E. M. Pothos and A. J. Wills (eds.) Formal Approaches in Categorization. Cambridge: CUP. 18–39.

Peirce, Jonathan, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman and Jonas Kristoffer Lindeløv. 2019. PsychoPy2: Experiments in behavior made easy. Behavior Research Methods 51(1). 195–203.

R Core Team. 2021. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Rácz, Péter. 2019. Frequency and prototypicality determine variation in the Hungarian verbal 1SG.INDEF. Acta Linguistica Academica 66(4). 601–620.

Rácz, Péter, Péter Rebrus and Miklós Törkenczy. 2021. Attractors of variation in Hungarian inflectional morphology. Corpus Linguistics and Linguistic Theory 17(2). 287–317.

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

Stan Development Team. 2019. RStan: The R interface to Stan. http://mc-stan.org/.

Trón, Viktor, Péter Halácsy, Péter Rebrus, András Rung, Péter Vajda and Eszter Simon. 2006. Morphdb.hu: Hungarian lexical database and morphological grammar. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). 1670–1673.