Article

Péter Rácz* and Péter Rebrus

# Lexical patterns in Hungarian vowel harmony

https://doi.org/10.1515/cllt-2024-0021
Received February 26, 2024; accepted November 15, 2024; published online December 13, 2024

**Abstract:** Hungarian shows variable front vowel harmony, particularly in suffixed back vowel + [ɛ] nouns. The study aims to address two main research questions: (1) To what extent does stem-level information (similarity across stems) predict suffix variation for back vowel + [ɛ] stems in Hungarian corpus data? (2) Do suffixes themselves predict suffix variation beyond the stem-level information? We draw on a dataset of 200 noun stems, 4,501 suffixed forms and $4 \times 10^6$ tokens, based on the New Hungarian Webcorpus, and use a K-Nearest Neighbours learner and a hierarchical generalised linear model to address these questions. We find that the majority of back vowel + [ɛ] stems show variable vowel harmony, that this depends on stem similarity and that similarity effects are amplified by vowel-initial suffixes. This points to a model of Hungarian vowel harmony in which stem- and suffix-level information are lexically specified.

**Keywords:** vowel harmony; Hungarian; learning models; mental lexicon

## 1 Background

Hungarian shows active vowel harmony: the last vowel of the stem regularly determines the suffix vowel. This paper is about back/front harmony, where a back stem vowel selects for a back suffix vowel and a front stem vowel selects for a front suffix vowel. Back/front harmony shows individual and lexical variation.

The Hungarian vowel system can be seen in Table 1, adapted from Siptár and Törkenczy (2000). There is an unrounded and a rounded set of high and low front vowels, while roundedness is not phonologically relevant for back vowels. All vowels

***Corresponding author: Péter Rácz**, Cognitive Science Department, Faculty of Natural Sciences, Budapest University of Technology and Economics, Budapest, Hungary,
E-mail: racz.peter.marton@ttk.bme.hu
**Péter Rebrus**, HUN-REN Hungarian Research Centre for Linguistics & Eötvös Loránd University, Budapest, Hungary

**Table 1:** Hungarian vowels.

| | Front | | | | Back | |
|---|---|---|---|---|---|---|
| | Unrounded | | Rounded | | | |
| High | i | iː | y | yː | u | uː |
| Mid | | eː | ø | øː | o | oː |
| Low | ɛ | | | | ɒ | aː |

have long-short pairs. The pairs [ɛ]/[eː] and [ɒ]/[aː] are different in both length and quality: [ɛ] is lower than [eː] and [aː] is lower than [ɒ] (the latter, phonetic, distinction is not marked in the table).

In traditional descriptions, front unrounded vowels, marked in grey in Table 1, are seen as neutral with respect to back/front harmony.

In practice, stems vary. Siptár and Törkenczy (2000) provide a typology of this variation. They distinguish four stem classes of interest. We adopted their system with slight changes, as seen in Table 2.

i. Simple harmonic stems consist of only front or back vowels, and select for a front or back suffix, respectively.

ii. Complex harmonic stems contain both front and back vowels, and suffix vowel is selected for by the final vowel of the stem.

iii. Simple neutral stems contain back vowels with a final neutral front vowel and also select for back suffixes.

iv. Complex neutral stems have back vowel(s) + front low unrounded [ɛ], which vacillates as a neutral vowel. (We ignore the behaviour of the other front vowels here.) The extent of this vacillation depends on stem type.

**Table 2:** Hungarian back/front harmony.

| Class | Subclass | Stem | Dative | Gloss |
|---|---|---|---|---|
| i. Simple harmonic | | kosoruː | kosoruː-nɒk | *wreath* |
| | | køsøryː | køsøryː-nɛk | *grinding wheel* |
| ii. Complex harmonic | | nyɒns | nyɒns-nɒk | *nuance* |
| | | ʃoføːr | ʃoføːr-nɛk | *chauffeur* |
| iii. Simple neutral | | pɒpiːr | pɒpiːr-nɒk | *paper* |
| | | kaːveː | kaːveː-nɒk | *coffee* |
| iv. Complex neutral | a. True neutral | hɒvɛr | hɒvɛr-nɒk | *pal* |
| | b. Disharmonic | koːdɛks | koːdɛks-nɛk | *codex* |
| | c. Vacillating | dʒungɛl | dʒungɛl-nɒk/nɛk | *jungle* |

(a) True neutral stems have back vowels + [ɛ] and predominantly select for a back suffix.
(b) Disharmonic stems have back vowels + [ɛ] and predominantly select for a front suffix.
(c) Vacillating stems have back vowels + [ɛ] and are widely attested with both back and front vowel suffixes.

Across- and within-stem variation in the complex neutral class is shaped by the phonological composition of the stem. As Hayes et al. (2009) find, back vowel + [ɛ] stems are more likely to prefer front suffixes if the stem ends in a bilabial stop, a sibilant, a coronal sonorant or a consonant cluster. Native speakers carry over these patterns to non-word stems in a Wug task (Berko 1958), leading Hayes et al. to stipulate that these consonantal triggers, and, broadly speaking, specific phonologically unnatural patterns, have to be represented in the speaker grammar.

Within-stem variation is driven by one additional factor, the suffix-initial vowel. Most suffixes exhibit back/front harmony: they agree with the last vowel of the stem. In addition, some suffixes are always consonant-initial, while others have a linking vowel with consonant-final stems. Some suffixes are always vowel-initial. This can be seen in the example of the dative, the plural and the causalis in Table 3.

- The dative has a front- and back-vowel alternant. It is always consonant-initial.
- The plural has a linking vowel if the stem ends in a consonant. The linking vowel is front or back, depending on the stem.
- The causalis always has a front vowel. It is always vowel-initial. (There exist vowel-initial suffixes, like the essive, that do harmonise with the stem.)

The presence and quality of the suffix-initial vowel can be predicted from the morphophonology of the stem and the suffix, to an extent (see Rebrus et al. 2024). Descriptions of Hungarian would make a difference between a linking vowel (present with a consonant-final stem) and a suffix-initial vowel. We only make a distinction between consonant-initial and vowel-initial suffixes. (Note that, as we will

**Table 3:** Stems and suffix-initial vowels.

| Stem type | Stem | Dative | Plural | Causalis | Gloss |
|---|---|---|---|---|---|
| C-final back | laːɲ | laːɲ-nɒk | laːɲ-ok | laːɲ-eːrt | *girl* |
| C-final front | leːɲ | leːɲ-nɛk | leːɲ-ɛk | leːɲ-eːrt | *being* |
| V-final back | holloː | holloː-nɒk | holloː-k | holloː-eːrt | *raven* |
| V-final front | hylløː | hylløː-nɛk | hylløː-k | hylløː-eːrt | *reptile* |

see, all the stems in our analysis are consonant-final, so linking vowels will always be present).

Rebrus et al. (2024) argue that the complex neutral stems under (iv) in Table 2 all show variation to some extent and that this partly depends on the suffix. According to this account, the true neutral class (iva) has examples of both back and front vowel suffixes. It is much more likely to select for a back vowel with a vowel-initial suffix, such that front [hɒvɛr-ɛk] (pal-pl) is much less likely than front [hɒvɛr-nɛk] (pal-dat). Conversely, the disharmonic class (ivb) also has examples of both back and front vowel suffixes. It is much more likely to select for a front vowel in a vowel-initial suffix, such that back [koːdɛks-ok] (codex-pl) is much less likely than back [koːdɛks-nɒk] (codex-dat).

Active vowel harmony in Hungarian has received considerable attention in the literature (Goldsmith 1985; Hayes et al. 2009; Kertész 2003; Rebrus et al. 2012; Siptár and Törkenczy 2000; Törkenczy 2011; Van der Hulst 2016; Zuraw and Hayes 2017).

Along with tonal systems, harmony systems have been one of the early successes of autosegmental phonology (see Goldsmith 1985), which worked with the assumption that the context of vowel-specific operations can be separate from the context of consonant-specific ones. The autosegmental framework can capture dependencies across vowels without regard to the intervening consonants, a major constraint on systems that use linear representations with one-to-one correspondence between segments and features.

As we have seen, more recent work on Hungarian vowel harmony, like Hayes et al. (2009) and Rebrus et al. (2024), is difficult to reconcile with the autosegmental account. This work suggests that the stem and the suffix together determine variable behaviour in back/front harmony and puts the Hungarian pattern squarely among morpho-phonological processes that make reference to lexical information (Hay and Baayen 2005; Lindsay-Smith et al. 2024).

Whether we assume a model of speaker production that incorporates lexical generalisations in a grammatical component or relies on analogy processes in the mental lexicon is a separate question. Current evidence suggests, however, that variable Hungarian vowel harmony makes reference to lexical information in some way (see Zuraw and Hayes 2017).

A number of open questions remain. To what extent is variable vowel harmony gradient or categorical? How broad is the distribution of stems that show within-stem variation? (Referencing Table 2, can we speak of categorical disharmonic and true neutral stems, or only vacillating stems?) How much and what type of lexical information is necessary to capture this distribution?

Answers to these questions have consequences for an adequate description of Hungarian variable vowel harmony in particular and a cross-linguistic typology of vowel harmony in general.

## 1.1 Present study

The aim of this study is to revisit front/back suffix vowel variation with bisyllabic back vowel + [ɛ] noun stems. We created a dataset of variable stems based on the new Hungarian Webcorpus (Nemeskey 2020).

We had two research questions:

1. To what extent does stem-level information (similarity across stems) predict suffix variation for back vowel + [ɛ] stems in Hungarian corpus data?
2. To what extent do suffixes themselves predict suffix variation above and beyond stems? That is, do we need to make reference to the stem or the entire form in an account of suffix variation?

Corpus data could support a number of possible scenarios. First, we have expectations on lexical **distributions**. It is possible that, within the complex neutral class (Table 2, iv.), the behaviour of true neutral (iva) and disharmonic (ivb) stems is near-categorical and the vacillating subclass (ivc) shows free variation. This would suggest that the whole complex neutral stem class is lexically idiosyncratic. Alternatively, the subclasses of the complex neutral class might be binning a continuous distribution of stems ranging from back- to front-suffix preference. This continuous distribution might result from free variation or be sensitive to stem composition. Second, we have expectations on the effects of **stem phonology** (Hayes et al. 2009) and **stem similarity** (Rebrus et al. 2024). Third, the effect of **suffix-initial vowels**, that is, vowel- versus consonant-initial suffixes on front variation might be either negligible, or present but unidirectional (such that vowel-initial suffixes always prefer back vowels, for instance), or else interact with stem type.

We used a K-Nearest Neighbours learner (see Peterson 2009) to check whether lexical similarity across stems predicts suffix variation. We used generalised linear models to test whether suffix-level information contributes over stem-level information in accounting for suffix variation.

# 2 Methods

## 2.1 Tools

We implemented the K Nearest-Neighbours model in R (R Core Team 2023). We used the packages *ggplot2* (Wickham 2011), *patchwork* (Pedersen 2024) and *sjPlot* (Lüdecke 2023) for visualisation, *lme4* (Bates et al. 2015), *performance* (Lüdecke et al. 2021) and *broom* (Robinson et al. 2023) for modelling.

## 2.2 Dataset

We compiled a frequency list from the Hungarian Webcorpus 2 (Nemeskey 2020). The Webcorpus contains $1.8 \times 10^7$ types and $8 \times 10^9$ tokens. We filtered the frequency list to include noun forms of two syllables with a back vowel + <e> (the regular spelling of [ɛ]). We used the *hunspell* spellchecker package (Ooms 2023) and hand-filtering to winnow the list. We picked the 30 most common harmonising suffix types that co-occur with these nouns. The resulting list has 200 stems and 4,501 suffixed forms, and a frequency count of $3.75 \times 10^6$ across all forms. All stems are consonant-final. For details, see the Supplementary Information.

We restricted the data to suffixed forms that do show back/front variation in the corpus, resulting in 164 stems and 2,462 suffixed forms. We went on to calculate the log odds ratio of back and front forms for each suffixed form $(\log \frac{\text{back}}{\text{front}})$, resulting in 1,231 suffixed pairs across 161 stems. The difference here arises because 3 stems do not show variation for any given suffixed form, only across suffixes.

A random sample of the resulting dataset can be seen in Table 4. The data were written. We will use transcriptions in the International Phonetic Alphabet to make the examples consistent and legible.

**Table 4:** Random sample of our dataset.

| Stem | Suffix | Form | Back | Front | $\log \frac{back}{front}$ | Gloss |
|---|---|---|---:|---:|---:|---|
| hɒvɛr | DAT | hɒvɛr-nɒk/hɒvɛr-nɛk | 3,259 | 161 | 3.01 | pal |
| notɛs | POSS.1SG | notɛs-om/notɛs-ɛm | 80 | 219 | −1.01 | pocketbook |
| puːdɛr | INE | puːdɛr-bɒ/puːdɛr-bɛ | 12 | 77 | −1.86 | makeup |
| fotɛl | PL | fotɛl-ok/fotɛl-ɛk | 754 | 5,282 | −1.95 | armchair |
| bɒlɛtː | ACC | bɒlɛtː-ot/bɒlɛtː-ɛt | 604 | 4,587 | −2.03 | ballet |
| lɒtɛks | ELA | lɒtɛks-roːl/lɒtɛks-røːl | 39 | 737 | −2.93 | latex |

Table 4 shows one variable stem + suffix pair per row. The columns are the stems attested in the corpus, the variable suffix, the variants, the number of back and front forms in the corpus, their log odds and the gloss for the stem.

## 2.3 Stem effects: K-nearest neighbour learner

To create our training data, we took the 164 varying stems and transcribed them using a simplified phonetic transcription. This transcription replaced letter digraphs with single characters (*szoftver* [softʋɛr] (‘software’) → <softver>). We calculated the log odds of back/front forms for each stem by grouping the data across stems and summing back and front counts across suffixes:

$$\mathrm{lo}_{\mathrm{stem}} = \log \frac{\sum \mathrm{back}}{\sum \mathrm{front}}$$

An alternative approach would have been to follow Janda et al. (2010) and fit a generalised linear mixed model on all forms with the odds of back/front variants as the outcome variable and estimate a random intercept for stems and for suffixes. We could then use the stem random intercept in place of the raw totals. However, the correlation between this random intercept and the raw log odds across stems is $r = 0.99$, and this makes little practical difference. We used stem frequency to split the stems into five frequency quantiles.

Our K-Nearest Neighbour learner was written in *R*. It matched a target word to training words and predicted its behaviour based on the behaviour of its nearest neighbours. It calculated the Levenshtein distance between the transcribed test word and transcribed training words, arranged training words from smallest to largest distance from the test word and selected the first k training words. Some training words might have the same Levenshtein distance from the test word (e.g. the Levenshtein distance between <hotel>, <motel> and <fotel> ('armchair') is 1), so the order of target words within distance brackets was randomised. The learner then summed over the back and front form counts for the k nearest neighbours and calculated a total log odds, using the formula:

$$\log \frac{\sum^k \mathrm{back}}{\sum^k \mathrm{front}}$$

where $\sum^k$ is the sum of counts for the first $k$ forms, *back* refers to forms with a back vowel suffix and *front* to those with a front vowel suffix. The learner returned this value as the prediction for the test form. The learner used a leave-one-out fitting method, comparing test forms to all training forms except the test form itself.

Our learner differs from the K-Nearest Neighbours (KNN) learner used in categorisation problems and machine learning. A more typical KNN learner provides a category label, not a category weight. In addition, a more typical KNN learner will not involve a random component, since distances in any given category space are likely more fine-grained and so unique for every target item in the training set. (E.g. on an RGB scale, every unit of change in R/G/B from a reference colour will define a distinct colour, however small the difference is.)

Our learner had two hyperparameters, $k$, the number of nearest neighbours (possible values: 1, 7, 10, 12, 15), and $f$, the relative frequency of stems in the training set (possible values: 1 – 5, where the training set consists of forms in the $f \leq$ quantiles of the total training set). Higher values of $k$ involve comparison of the target stem to more neighbours, while higher values of *f* mean that the comparison set will be increasingly limited to higher frequency forms.

For each hyperparameter setting, we fit a binomial generalised linear model predicting the back/front ratio for each stem from the KNN prediction for that stem across all 164 stems. We used the linear model's *z*-value to select the best model. Since models only differed from one another by the KNN hyperparameter settings, this gave us the best KNN hyperparameters: $k = 7$ and $f = 3$. This means that the best learner compared the target form to its first seven nearest neighbours. The best learner operated on the top 40 % of the frequency distribution of training forms, ignoring the less frequent training forms.

A K Nearest-Neighbour Learner is very likely not the state-of-the-art model of lexical similarity, but it is sufficient to demonstrate the relevance of stem-level similarity for suffix variation. We return to this briefly in the Discussion. In addition, a distance metric that incorporates segmental similarity (where voiceless coronal stop <t> will be more similar to voiced coronal stop <d> than to voiced labial stop <b>) will be likely more accurate in expressing the role of similarity in a learning model. At the same time, Rácz et al. (2024) have explored various distance metrics in modelling Hungarian morphophonological variation, including vowel harmony, and found that models based on Levenshtein distance have comparable accuracy to models using segmental similarity.

## 2.4 Suffix effects: generalised linear models

The K-Nearest Neigbours learner makes predictions for stems only. In order to incorporate suffix-specific information on some level, we marked whether a suffix was **vowel-** or **consonant-initial** in our paired dataset. We then went on to build generalised linear mixed models that used stem-level information (learner predictions), and suffix-level information (whether the suffix is consonant- or vowel-initial), to predict the log odds of variable forms in the data using the formula:

$$\frac{\text{back}}{\text{front}} \sim \text{learner} \quad \text{weight} \ + \ \text{suffix} - \text{initial} \quad \text{vowel}$$

where learner weight is the back/front ratio of the stem as predicted by the best KNN learner and suffix-initial vowel specifies whether the suffix is vowel-initial or consonant-initial. We built four models, seen in Table 5. In establishing the random effect structure, we followed Janda et al. (2010). We used AIC, BIC and a likelihood ratio test of model fit to find the best random effect structure for each model and to find the best model.

**Table 5:** Generalised linear mixed models fit on the data: formula, statistics (AIC, BIC, conditional and marginal $R^2$, RMSE), pairwise likelihood tests (1 vs. 2, 2 vs. 3, 3 vs. 4): $\chi^2$ difference and *p* value.

| *n* | Formula | AIC | BIC | R2 c. | R2 m. | RMSE | $\chi^2$ diff. | *p* |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 21,115.72 | 21,131.07 | 0.68 | 0.00 | 0.15 | | |
| 2 | 1 + knn | 21,095.15 | 21,115.62 | 0.68 | 0.08 | 0.15 | 22.57 | <0.001 |
| 3 | 1 + knn + suffix-initial vowel | 16,574.85 | 16,610.66 | 0.69 | 0.07 | 0.14 | 4,526.31 | <0.001 |
| 4 | 1 + knn * suffix-initial vowel | 16,563.04 | 16,603.96 | 0.70 | 0.11 | 0.14 | 13.81 | <0.001 |

Table 5 shows the four models fit on the data along with their AIC, BIC, $R^2$ and root mean square error (RMSE). The model also shows the results of pairwise likelihood ratio tests (model 1 vs. 2, 2 vs. 3, 3 vs. 4). Model 1 only estimates an intercept and a random intercept for stems and suffixes. Model 2 adds the weight that is estimated by the K Nearest-Neighbour learner. As we see in the AIC and BIC scores that Model 2 is a better fit than Model 1 – the learner weights explain some variation in the data. Model 3 also includes whether the suffix is vowel- or consonant-initial, that is, whether a suffix-initial vowel is present or absent in the suffixed form. It gives a much better fit than Model 2 – the presence of the suffix-initial vowel is hugely informative in determining the suffix vowel. Model 4 includes an interaction of learner weight and suffix-initial vowel. This model provides the best fit.

# 3 Results

We start the section by reporting on the raw corpus data and then build up the best model.

## 3.1 Distributions in the corpus

In our sample, 200 stems fit the template for the complex neutral class (iv) in Table 2. A total of 164 stems vary across all suffix types. Distributions can be seen in Figure 1. When we consider stems separately with vowel- and consonant-initial suffixes, we find that the majority of stems show vacillating behaviour in both sets, and more stems show disharmonic behaviour (categorical preference for the front suffix) with vowel-initial suffixes (top panel). When we consider back-suffix preference for vacillating stems only, these tend toward disharmonic behaviour (preference for a front suffix), and this preference is more apparent with vowel-initial suffixes (bottom panel). Note that, even among vacillating stems, some show extreme preference for

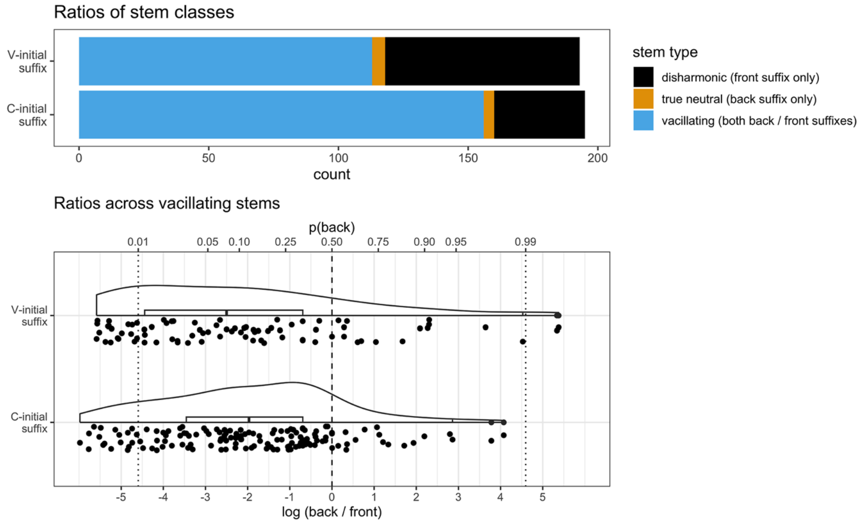


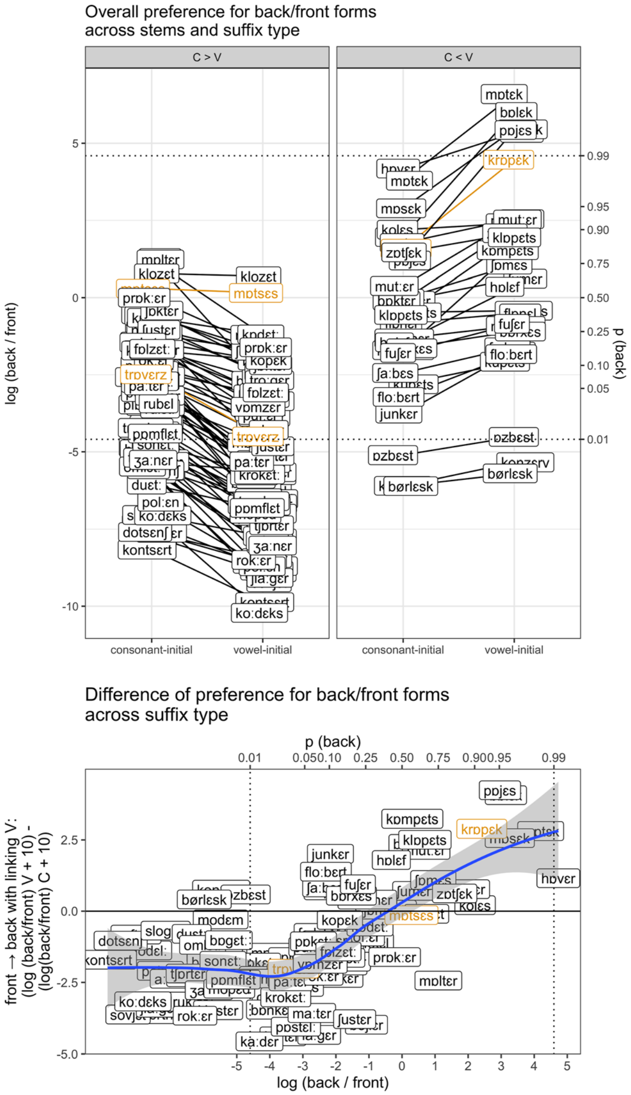
**Figure 1:** Counts of true neutral (only back suffixes), disharmonic (only front suffixes) and vacillating (both back and front suffixes) stems in the corpus (top panel). Distributions of vacillating stems for back-suffix preference, across vowel-initial and consonant-initial suffixes. Scale in log odds back (bottom)/p (back) (top). The dotted vertical lines show 1 % and 99 % preference, the dashed line, 50 % preference, for the back suffix (bottom panel).

back or front suffixes. Stems left of the left-hand dotted line have at least a 99:1 ratio of front suffixes. Stems right of the right-hand dotted line have at least a 99:1 ratio of back suffixes. Interesting variation happens in between.

The class-based terminology is restrictive in the sense that, among variable stems, the same stem might show more ‘disharmonic’ or ‘neutral’ behaviour depending on the presence of the suffix-initial vowel. This can be seen in Figure 2.

The top panel of the figure is the bottom panel of Figure 1, rotated 270 degrees. For each stem, we see the back suffix preference across suffixes with and without a suffix-initial vowel. When stems have a strong neutral tendency (top right panel), this is more pronounced with vowel-initial suffixes. If stems have a disharmonic tendency (top left panel), then this is more pronounced with vowel-initial suffixes. There are two important caveats: first, vacillating stems are, by definition, variable. Second, below a log odds of –5, a stem will show front preference 99.33 % of the time, so differences across suffix-initial vowel in this tail of the distribution are actually negligible.

We see the magnitude of difference between vowel- and consonant-initial suffixes across magnitude of overall difference for each stem in the bottom panel


Overall preference for back/front forms across stems and suffix type

Difference of preference for back/front forms across suffix type

**Figure 2:** Back suffix preference for each vacillating stem across consonant- and vowel-initial suffixes, split into two panels based on overall preference (top panel). The absolute difference of back-preference with vowel-initial versus consonant-initial suffixes compared to overall rate of back-preference (bottom panel). Coloured examples are discussed in the running text.

of Figure 2. If a stem shows meaningful vacillating behaviour, the difference based on the suffix-initial vowel will steadily increase as a function of back-preference.

A few examples will put this relationship into context. For the word *traverz* [trɒvɛrz] 'beam', 3 % of all suffixed forms have a back suffix. This increases to 7 % (more than double) for consonant-initial suffixes only. It decreases to 1 % (one third) for vowel-initial suffixes only. The low overall preference goes with a much lower preference when a suffix-initial vowel is present: [trɒvɛrz-nɒk] 'beam-dat' is more likely than [trɒvɛrz-ok] 'beam-pl'. For the word *krapek* [krɒpɛk] 'lad', the overall rate of back suffixes is 92 %. This goes *down* to 83 % with consonant-initial suffixes only and *up* to 99 % with vowel-initial suffixes only. The high overall preference goes with a much higher preference when a suffix-initial vowel is present: [krɒpɛk-ok] 'lad-pl' is more likely than [krɒpɛk-nɒk] 'lad-dat'. Compare these two words to a word in the middle of the overall distribution, *macesz* [mɒtsɛs] 'matzo': 57 % overall, 57 % with a suffix-initial consonant, 54 % with a suffix-initial vowel – roughly the same numbers. The three examples here are drawn from a clearly increasing trend: The suffix-initial vowel amplifies disharmonic/neutral tendencies across stems.

## 3.2 K nearest-neighbour predictions

In order to understand the factors underlying the shifting effect of the suffix-initial vowel, we need to first understand stem behaviour in itself. Is there a particular reason for [trɒvɛrz] to have a disharmonic preference and for [krɒpɛk] to have a neutral preference? (So that we usually find [trɒvɛrz-ɛk] 'beam-pl' and [krɒpɛk-ok] 'lad-pl'.)

The predictions of the K Nearest-Neighbours Model indicate that stem similarity drives stem behaviour, as seen in Figure 3. The main interpretation of this figure is that the more a stem looks like other stems that have a back vowel preference (vertical axis), the more the stem itself will have a back vowel preference (horizontal axis). For reference, the Pearson correlation between KNN similarity weights (how much the stem looks like other stems) and the log odds of back preference is $r = 0.34$. For a number of outliers, KNN weight clearly overestimates back preference.

## 3.3 Stem phonology

A flip side of stem similarity is stem phonology. Hayes et al. (2009) argue that a Hungarian complex neutral stem is more likely to select a front vowel suffix if it ends in a labial stop, a sibilant, a coronal sonorant or a consonant cluster. 142/164 stems match one of these categories in our data. We fit a generalised linear model predicting the log odds of back suffixes if the stem matched any of the categories.

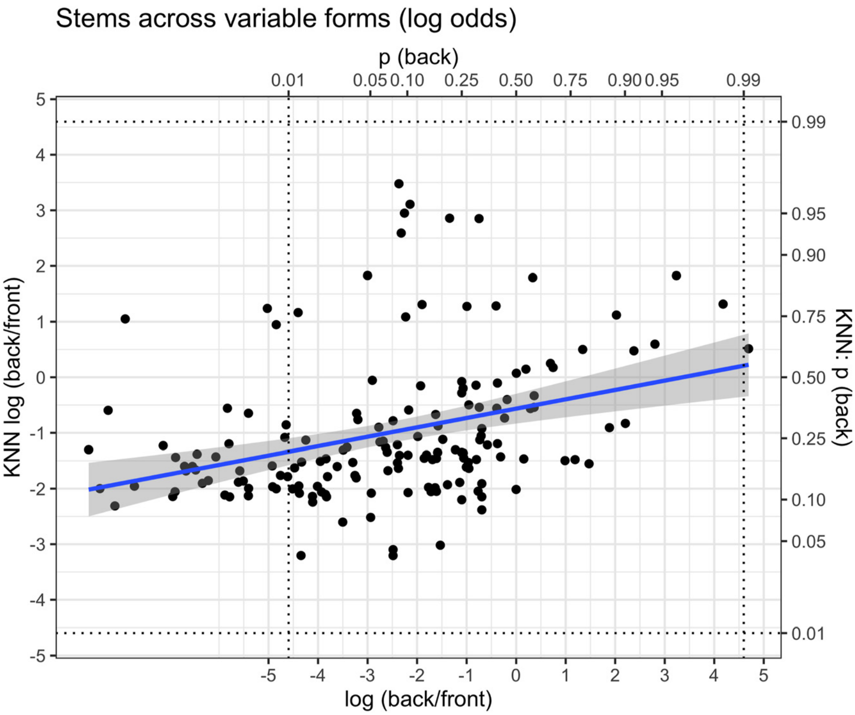


**Figure 3:** The log odds of back/front variants per stem (horizontal axis) and best KNN predictions for the stem (vertical axis).

We found that these stems are less likely to select a back suffix (est = –2, 95 % CI: [–1.99; 2.02]). This means that our data broadly replicate their findings.

## 3.4 Combined model predictions

Table 5 shows that the best model of back/front variation across forms in the data includes an interaction term of K Nearest-Neighbour learner weight and the suffix-initial vowel. That is, stem-level similarity and the presence of a suffix-initial vowel *together* predict back/front ratios across forms in the data. This can be seen in Figure 4, a prediction plot of the best model.

The vertical axis shows the overall model prediction for the log odds of back/front variants per stem-suffix pair. The horizontal axis shows the effect of the KNN learner prediction only: our best estimate of the effect of stem similarity. We see two trajectories, for stems with vowel-initial suffixes and the same stems with consonant-initial suffixes.

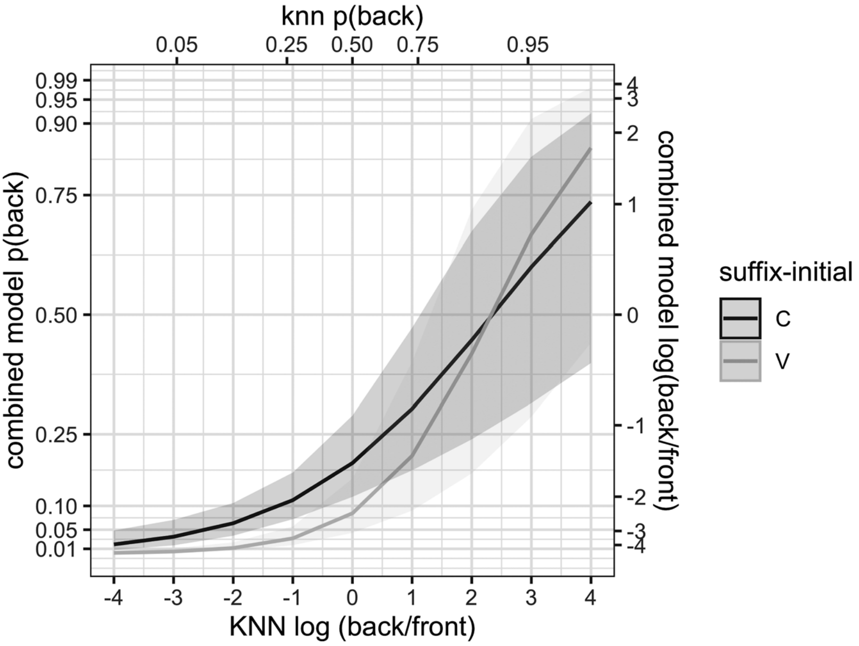


**Figure 4:** The log odds of back/front variants per stem (horizontal axis) and best KNN predictions for the stem (vertical axis).

If we look at forms with a suffix-initial vowel, the stem-similarity effect is stronger. As seen in Table 5, this difference is significant ($p$ < 0.001).

# 4 Discussion

We compiled a dataset of bisyllabic back vowel + [ɛ] stems from the new Hungarian Webcorpus and used a simple algorithmic learning model and hierarchical generalised logistic regression to test existing claims in the literature. We found that the majority of complex neutral stems are vacillating in our data. Stems behave like similar stems in their disharmonic/neutral preference (preference for a front or back suffix). This similarity effect is stronger when a suffix begins with a vowel.

When we look at stem and form distributions, we find a larger disharmonic class (stems that always take a front suffix) and a smaller true neutral class (stems that always take a back suffix). The vacillating stems in between are the largest group. They show a unimodal distribution in their behaviour, from mostly front suffixes (close to disharmonic) to mostly back suffixes (close to true neutral).

For a large number of stems, we can just about detect variability, with a handful of back/front suffixed forms in the data. This is despite the size of our dataset, the New Hungarian Webcorpus, which is about an order of magnitude larger than the previous Hungarian webcorpus (Halácsy et al. 2004; Nemeskey 2020). To us, this suggests that the appearance of categorical disharmonic and true neutral stems results from data scarcity – for these stems, the vacillating variants might exist but be very difficult to find in the wild. The reverse scenario, where most stems inherently vary but some do not, and the data reflect this, seems far less likely to us. This would mean that all of the complex neutral stems in Siptár and Törkenczy (2000) belong to the vacillating class, some with stronger disharmonic or neutral tendencies.

Stem similarity clearly plays a role in shaping stem preference. We know that such stem-level effects generalise to novel forms in Wug tasks (Hayes et al. 2009). Curiously, stem-level effects are amplified when the suffix begins with a vowel.

While this is primarily a descriptive paper, we want to speculate on the origin and the larger context of the observed patterns.

## 4.1 Diachronic context

Rebrus et al. (2022) note a semantic pattern underlying complex neutral stem behaviour, namely, that learned borrowings show a disharmonic tendency, while familiar borrowings show a neutral tendency. This is replicated in our data, as shown by Table 6. The table shows the ten most neutral and ten most disharmonic variable stems in our dataset, with the strongest and weakest back suffix preference, respectively. The most neutral examples are all informal, familiar words, colloquialisms or diminutives, while the most disharmonic examples are all learned borrowings.

Forró (2013) and Rebrus et al. (2022) suggest that colloquial and learned forms were borrowed with different phonemic realisations for the [ɛ]. The phonetic differences, in time, disappeared, while the resulting suffixation preferences remained.

**Table 6:** The most neutral and most disharmonic stems in our data.

| Back preference rank | Stem | Gloss |
|---|---|---|
| Most neutral | hɒvɛr, mɒtɛk, fɒtɛr, mɒsɛk, bɒlɛk, pɒjɛs, krɒpɛk, kolɛs, mutɛr, komplɛtː | pal, maths, dad, tradie, dupe, sideburns, fellow, dorm, mum, complete |
| Most disharmonic | maːgnɛʃ, goːlɛm, modɛlː, konsɛrn, koːdɛks, softvɛr, projɛkt, sovjɛt, dotsɛnʃ, koncɛrt | magnet, golem, model, firm, codex, software, project, soviet, lecturer, concert |

This tentative account would need to be cross-referenced with diachronic and dialectal data. What remains true is that source of borrowing and variable behaviour are correlated for the complex neutral stem class.

If we overlay the source of borrowings in our data on the front/back suffix vowel distribution, we see clear clusters of older and more recent borrowings. This can be seen in Figure 5, which shows the disharmonic/neutral preferences of variable stems (like in the bottom panel of Figure 1), across the five main source languages of the borrowings (Gerstner et al. 2024; Zaicz et al. 2006). Words with Yiddish and German origin, which are overwhelmingly recent, colloquial borrowings, are markedly more neutral than the rest. Words of French and Latin origin, which tend to be older, learned borrowings, are more disharmonic than the other groups. A Yiddish borrowing is far more likely to have a back suffix than a Latin borrowing in our data.

This account would presuppose that suffix vowel preference is gradient and lexically specified. The K-Nearest Neighbour model is likely not the most accurate or cognitively plausible model of lexically grounded suffix selection. More complex algorithmic learners, like the Generalised Context Model (Nosofsky 2011), have seen considerable success in modelling morphologically complex languages in general

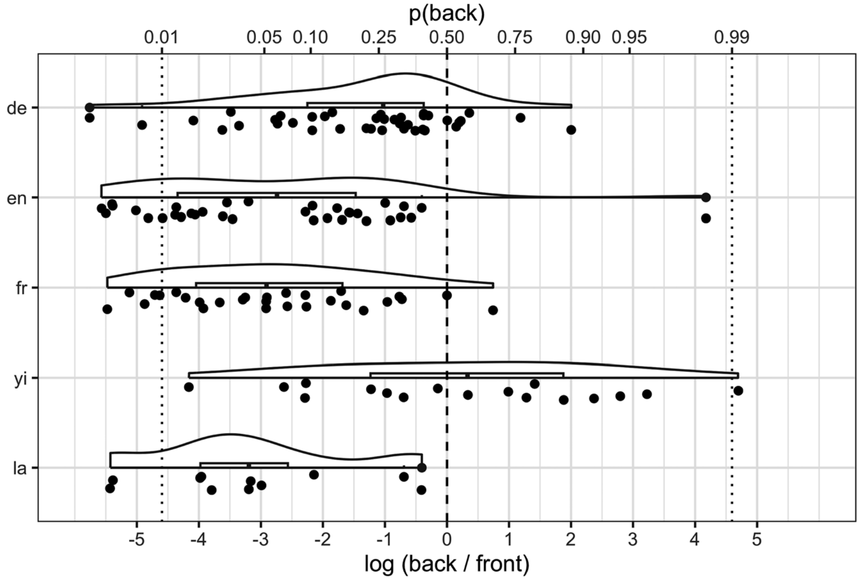


**Figure 5:** Distributions of vacillating stems for back-suffix preference, from disharmonic (all front) to true neutral (all back), across source language (German (de), English (en), French (fr), Yiddish (yi) and Latin (la)). Scale in log odds back (bottom)/p (back) (top). The dotted vertical lines show 1 % and 99 % preference, the dashed line, 50 % preference, for the back suffix.

(see e.g. Dawdy-Hesterberg and Pierrehumbert 2014) and Hungarian in particular (see e.g. Rácz et al. 2021). The best fit might even be provided by a model that captures stem-level generalisations without any explicit reference to the words in the mental lexicon, like Naive Discriminative Learning (Baayen 2010), Minimal Generalisation (Albright and Hayes 2003) or Harmonic Grammar (Zuraw and Hayes 2017).

## 4.2 Paradigmatic context

Whatever the source of stem-level effects, these are clearly amplified by a suffix-initial vowel, as seen in Figures 2 and 4. One possible underlying factor is the discrepant frequency distribution of vowel-initial and consonant-initial suffixes, which can be seen in Figure 6. First, 60 % of all suffixed forms in our data are vowel-initial. What is more, 50 % of vowel-initial suffixed forms and 30 % of all suffixed forms belong to the most frequent Hungarian noun suffix, the plural, which has a linking vowel with consonant-final stems. Two things to note here: first, the same is true for the second most frequent noun suffix, the accusative. Second, all our stems are consonant-final.

We can tie together the frequency distribution of suffixes in Figure 6 and the behaviour of suffixed forms with vowel-initial versus consonant-initial suffixes if we assume that not only stems but also suffixes are represented in the mental lexicon. If a form has a suffix-initial vowel, the suffix is very often the same, the plural. If a form has no suffix-initial vowel, the suffix can be any one of a number of different consonant-initial suffixes. In a hierarchical, richly detailed lexicon (Johnson 2006), the weights that determine behaviour for an individual form will be partially pooled

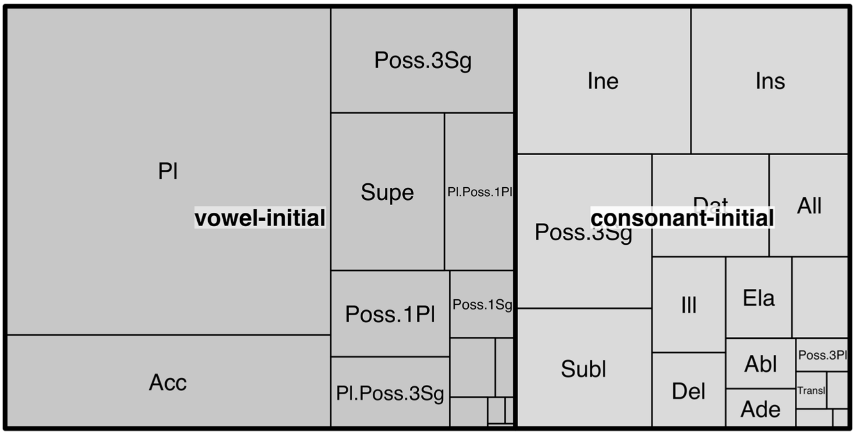


**Figure 6:** Tree map of suffix frequencies across all data. Rectangle size shows frequency.

over both stems and suffixes (as in a hierarchical model, see Bates et al. 2015). Stem preference will have a stronger effect for forms with a suffix-initial vowel, which have a simpler hierarchical structure, as compared to forms without a suffix-initial vowel, which have more competing pressures from a range of suffixes. This would point toward a paradigm-based, as opposed to simply harmonic, morphology of Hungarian, as suggested by Rebrus et al. (2024).

In summary, the behaviour of Hungarian complex neutral noun stems is gradient, rather than categorical. It is shaped by the back suffix vowel preference of the stem, which, in turn, is driven by similarity and has a tentative historical explanation. It is also shaped by the presence or absence of a suffix-initial vowel, which is evidence for a paradigm-based, rather than concatenating – autosegmental model of vowel harmony.

## 4.3 Future directions

Our analysis uncovers a number of interesting patterns in Hungarian corpus data, which support a lexical account of variation in vowel harmony. It does not tell the whole story. In particular, we see three ways of expanding this analysis.

First, we briefly reference but do not elaborate on the role of stem phonology, highlighted by Hayes et al. (2009), in Section 3.3. One could systematically explore the strength of the specific phonological constraints proposed by Hayes and colleagues (does the stem end in a labial stop, a sibilant, etc.) and compare them to a measure of analogical similarity in a single model that allows for regularisation over collinear predictors (Emmert-Streib and Dehmer 2019). This could show if our lexical effects can be completely reduced to more parsimonious abstract generalisations.

Second, both our proposed diachronic and paradigmatic explanation remain somewhat circular if these are only applied to corpus data. A production experiment could evaluate whether (a) non-word targets act like neologisms and pattern with more recent borrowings and (b) show any paradigmatic effects – we would not trivially expect this from non-words which have no attested plural in the ambient language.

Third, if we explore these two facets of this variation, we would be in a better position to link the pattern we discuss in this paper to broader theories of language variation and change (see Hayes 2022; Pierrehumbert 2016).

## 4.4 Conclusions

Accounts of Hungarian vowel harmony have pointed at stem-level effects (Rebrus et al. 2022), and, to our knowledge, ours is the first study to use algorithmic learning over corpus data to support these arguments. In addition, accounts of vowel harmony in general either assume no lexical effects (Van der Hulst 2016) or effects restricted to stem-control (Hayes et al. 2009), assuming that all harmonically alternating suffixes harmonise in the same way when combined with the same type of stem. Our results show a suffix effect, offering an update of the broader understanding of the typology of vowel harmony systems.

## Supplementary Material

The supplementary material for this article is available at https://doi.org/10.5281/zenodo.12699305.

**Acknowledgments:** We are grateful to our Editor, our Reviewers and Vásárhelyi Dani.
**Conflicts of interest:** The authors declare no conflict of interest.
**Research funding:** This research was supported by grants NKFI-K139271, NKFI-FK138188, TKP2021-EGA-02, OTKA-138188 and OTKA-139271 by the Hungarian National Research, Development and Innovation Fund. This was also supported by grant TKP2023-NVA-2 by Innovációs és Technológiai Minisztérium.

## References

Albright, Adam & Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90(2). 119–161.

Baayen, R. Harald. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon* 5(3). 436–461.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.

Berko, Jean. 1958. The child's learning of English morphology. *Word* 14(2–3). 150–177.

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

Emmert-Streib, Frank & Matthias Dehmer. 2019. High-dimensional lasso-based computational regression models: Regularization, shrinkage, and selection. *Machine Learning and Knowledge Extraction* 1(1). 359–383.

Forró, Orsolya. 2013. *Ingadozás a magyar elölségi harmóniában [variation in Hungarian backness harmony]*. Pázmány Péter Katolikus Egyetem Dissertation.

Gerstner, Károly, Zita Horváth-Papp, Andrea Kacskovics-Reményi, László Horváth, Zsuzsanna Molnár, Mária Hochbauer, Dóra Tamás, Attila Mártonfi & Csaba Merényi. 2024. UESzWeb Új Magyar Etimológiai Szótár — uesz.nytud.hu. https://uesz.nytud.hu/index.html (accessed 5 June 2024).

Goldsmith, John. 1985. Vowel harmony in Khalkha Mongolian, Yaka, Finnish and Hungarian. *Phonology* 2. 253–275.

Halácsy, Péter, András Kornai, Németh László, Rung András, István Szakadát & Trón Viktor. 2004. Creating open language resources for Hungarian. In *Proceedings of the 4th international conference on language resources and evaluation (LREC2004)*.

Hay, Jennifer B. & R. Harald Baayen. 2005. Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences* 9(7). 342–348.

Hayes, Bruce. 2022. Deriving the wug-shaped curve: A criterion for assessing formal theories of linguistic variation. *Annual Review of Linguistics* 8(1). 473–494.

Hayes, Bruce, Péter Siptár, Kie Zuraw & Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85. 822–863.

Van der Hulst, Harry. 2016. Vowel harmony. In *Oxford research encyclopedia of linguistics*. Oxford: OUP.

Janda, Laura A., Tore Nesset & 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.

Johnson, Keith. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics* 34(4). 485–499.

Kertész, Zsuzsa. 2003. Vowel harmony and the stratified lexicon of Hungarian. *The Odd Yearbook* 7. 62–77.

Lindsay-Smith, Emily, Matthew Baerman, Sacha Beniamine, Helen Sims-Williams & Erich R. Round. 2024. Analogy in inflection. *Annual Review of Linguistics* 10. 211–231.

Lüdecke, Daniel. 2023. sjplot: Data visualization for statistics in social science. Available at: https://CRAN.R-project.org/package=sjPlot.Rpackageversion2.8.15.

Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner & Dominique Makowski. 2021. performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software* 6(60). 3139.

Nemeskey, Dávid Márk. 2020. *Natural language processing methods for language modeling*. Eötvös Loránd University PhD thesis.

Nosofsky, Robert M. 2011. The generalized context model: An exemplar model of classification. In Emmanuel M. Pothos & Andy J. Wills (eds.), *Formal approaches in categorization*, 18–39. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511921322.002.

Ooms, Jeroen. 2023. hunspell: High-performance stemmer, tokenizer, and spell checker. Available at: https://CRAN.R-project.org/package=hunspell.Rpackageversion3.0.3.

Pedersen, Thomas Lin. 2024. patchwork: The composer of plots. Available at: https://CRAN.R-project.org/package=patchwork.Rpackageversion1.2.0.

Peterson, Leif E. 2009. K-nearest neighbor. *Scholarpedia* 4(2). 1883.

Pierrehumbert, Janet B. 2016. Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics* 2. 33–52.

R Core Team. 2023. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

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

Rácz, Péter, Péter Rebrus & Szilárd Tóth. 2024. Evaluating an ensemble model of linguistic categorization on three variable morphological patterns in Hungarian. In *Proceedings of the annual meeting of the cognitive science society*, vol. 46.

Rebrus, Péter, Péter Szigetvári & Miklós Törkenczy. 2012. Dark secrets of Hungarian vowel harmony. In Eugeniusz Cyran, Henryk Kardela & Bogdan Szymanek (eds.), *Sound, structure and sense: Studies in memory of Edmund Gussmann*, 491–508. Lublin: Wydawnictwo KUL.

Rebrus, Péter, Péter Szigetvári & Miklós Törkenczy. 2022. How morphological is Hungarian vowel harmony? In *Proceedings of the annual meetings on phonology*.

Rebrus, Péter, Péter Szigetvári & Miklós Törkenczy. 2024. No lowering, only paradigms: A paradigm-based account of linking vowels in Hungarian. *Acta Linguistica Academica* 71(1–2). 137–170.

Robinson, David, Alex Hayes & Simon Couch. 2023. broom: Convert statistical objects into tidy tibbles. Available at: https://CRAN.R-project.org/package=broom.Rpackageversion1.0.5.

Siptár, Péter & Miklós Törkenczy. 2000. *The phonology of Hungarian*. Oxford, UK: OUP Oxford.

Törkenczy, Miklós. 2011. Hungarian vowel harmony. *The Blackwell Companion to Phonology* 5. 2963–2990.

Wickham, Hadley. 2011. ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics* 3(2). 180–185.

Zaicz, Gábor, Ildikó Tamás & Magda T. Somogyi. 2006. *Etimológiai szótár: Magyar szavak és toldalékok eredete*. Budapest: Tinta.

Zuraw, Kie & Bruce Hayes. 2017. Intersecting constraint families: An argument for harmonic grammar. *Language* 93. 497–548.